UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

Willian Farias Carvalho Oliveira

Estimating Human Age Using Machine Learning on Panoramic Radiographs for Multi-Regional Brazilian Patients

Recife

2024

Willian Farias Carvalho Oliveira

Estimating Human Age Using Machine Learning on Panoramic Radiographs for Multi-Regional Brazilian Patients

Dissertation presented to the Post Graduate Program in Computer Science at Centro de Informática, Universidade Federal de Pernambuco, as a partial requirement for obtaining the degree of Master in Computer Science.

Concentration Area: Computational Intelligence

Advisor: Cleber Zanchettin

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Oliveira, Willian Farias Carvalho.
Estimating human age using machine learning on panoramic radiographs for multi-regional brazilian patients / Willian Farias Carvalho Oliveira. - Recife, 2024.
67 f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciências da Computação, 2024.
Orientação: Cleber Zanchettin.
Inclui referências.

1. Ciências forense; 2. Rede neural profunda; 3. Estimativa de idade; 4. Métodos radiológicos. I. Zanchettin, Cleber. II. Título.

UFPE-Biblioteca Central

Willian Farias Carvalho Oliveira

“Estimating Human Age Using Machine Learning on Panoramic Radiographs for Multi-Regional Brazilian Patients”

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 10/12/2024.

BANCA EXAMINADORA

_____

Prof. Dr. Tsang Ing Ren
Centro de Informática / UFPE

_____

Profa. Dra. Deborah Queiroz de Freitas França
Departamento de Diagnóstico Oral / UNICAMP

_____

Prof. Dr. Cleber Zanchettin
Centro de Informática / UFPE
(**orientador**)

ACKNOWLEDGEMENTS

FACEPE and FAPESP funded the project's infrastructure through the Public Joint Call FAPESP-FACEPE 08/2022 – Support for Research in Applied Artificial Intelligence (AI). I thank both foundations for their support and contribution to the development of this research.

RESUMO

Este estudo aborda o desafio de desenvolver modelos de aprendizado de máquina para estimativa de idade a partir de radiografias panorâmicas de pacientes de diferentes regiões do Brasil. Utilizando dois conjuntos de dados geograficamente diversos — um da UFPE (Nordeste) e outro da Unicamp (Sudeste) — investigamos as limitações dos modelos de inteligência artificial quanto à predição de idade em contextos regionais distintos. Construímos um protocolo experimental para avaliar o comportamento dos modelos de aprendizagem de máquina em diferentes cenários. No primeiro experimento, o modelo treinado exclusivamente com dados do UFPE apresentou limitações ao ser testado em pacientes do Unicamp, resultando em um erro absoluto médio (MAE) de 3,10 anos no *dataset* UFPE e 4,97 anos no *dataset* Unicamp, evidenciando desafios de generalização. No segundo experimento, foram exploradas abordagens de *fine-tuning*, que, embora tenham melhorado o desempenho do modelo em dados regionais, não eliminaram totalmente os vieses. O terceiro experimento, treinando o modelo do zero em um conjunto de dados misto, alcançou o melhor equilíbrio entre precisão e generalização, com um MAE de 3,25 anos para o UFPE e 3,69 anos para o Unicamp, indicando maior robustez em relação às abordagens anteriores. O quarto experimento introduziu técnicas de aumento de dados para aprimorar a robustez do modelo contra outliers e casos extremos. Apesar de melhorias marginais, erros de alta magnitude persistiram, sugerindo a necessidade de estratégias adicionais, como técnicas mais avançadas de aumento de dados e arquiteturas mais complexas. Os resultados deste estudo reforçam a importância de conjuntos de dados diversificados e protocolos experimentais rigorosos para lidar com variabilidades regionais e características demográficas distintas. O modelo treinado em um conjunto de dados misto demonstrou ser a abordagem mais eficaz, destacando que a integração de populações diversas é crucial para aumentar a generalização dos modelos. Assim, o estudo contribui com evidências concretas para o desenvolvimento de sistemas mais robustos, capazes de serem aplicados de forma confiável em cenários clínicos e forenses.

Palavras-chave: Ciências Forenses, Estimativa de Idade, Rede Neural Profunda, Métodos Radiológicos.

ABSTRACT

This study addresses the challenge of developing machine learning models for age estimation based on panoramic radiographs of patients from different regions of Brazil. Using two geographically diverse datasets — one from UFPE (Northeast) and another from Unicamp (Southeast) — we investigated the limitations of artificial intelligence models in predicting age across distinct regional contexts. We designed an experimental protocol to evaluate the behavior of machine learning models in various scenarios. In the first experiment, the model trained exclusively on UFPE data showed limitations when tested on Unicamp patients, resulting in a mean absolute error (MAE) of 3.10 years on the UFPE dataset and 4.97 years on the Unicamp dataset, highlighting challenges in generalization.In the second experiment, fine-tuning approaches were explored, which, while improving the model's performance on regional data, did not completely eliminate biases. In the third experiment, training the model from scratch on a mixed dataset achieved the best balance between accuracy and generalization, with an MAE of 3.25 years for UFPE and 3.69 years for Unicamp, indicating greater robustness compared to previous approaches. The fourth experiment introduced data augmentation techniques to enhance the model's robustness against outliers and extreme cases. Despite marginal improvements, high-magnitude errors persisted, suggesting the need for additional strategies, such as more advanced data augmentation techniques and more complex architectures. The results of this study reinforce the importance of diverse datasets and rigorous experimental protocols to address regional variability and distinct demographic characteristics. The model trained on a mixed dataset proved to be the most effective approach, emphasizing that integrating diverse populations is crucial to improving model generalization. Thus, the study provides concrete evidence for the development of more robust systems capable of being reliably applied in clinical and forensic scenarios.

Keywords: Forensic Sciences, Age Estimation, Deep Neural Network, Radiological Methods.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Accurate age estimation plays a vital role in forensic science and civil investigations, contributing significantly to the reconstruction of biological profiles in missing-person cases, verification of the age of juvenile offenders, and situations where official identification documents are lacking. Conventional methods for age estimation often involve morphological, biochemical, and radiological analyses, with panoramic radiography standing out as the preferred approach due to its non-invasive nature, simplicity, and affordability. This technique allows for the comprehensive evaluation of dental development stages across all teeth simultaneously, thus establishing itself as an essential tool for age estimation (DALITZ, 1962; MOORREES; FANNING; JR, 1963; BANG; RAMM, 1970; DEMIRJIAN; GOLDSTEIN; TANNER, 1973; CAMERIERE; CINGOLANI; FERRANTE, 2004; SPALDING et al., 2005; ALKASS et al., 2010; RAJKUMARI et al., 2013; ELFAWAL; ALQATTAN; GHALLAB, 2015; BEKAERT et al., 2015; PURANIK; PRIYADARSHINI; UMA, 2015; CHEN et al., 2016; BENJAVONGKULCHAI; PITTAYAPAT, 2018; MÁRQUEZ-RUIZ et al., 2020; NOLLA et al., 1952).

Despite advances in radiological techniques, age estimation poses significant challenges, particularly in older individuals. Once dental development is completed, typically by the age of 24 with the closure of the third molar's apex, traditional manual and visual assessment methods become less effective, creating a gap in accurately determining age in later stages of life (MOORREES; FANNING; JR, 1963; DEMIRJIAN; GOLDSTEIN; TANNER, 1973). Methods like the pulp/tooth area ratio calculation have been explored to address aging in older individuals, focusing on the deposition of secondary dentin (CAMERIERE et al., 2006; CAMERIERE; FERRANTE; CINGOLANI, 2006; CAMERIERE et al., 2012). However, these methods also face limitations, including introducing bias from examiner subjectivity and decreased effectiveness after age 24 (MORSE et al., 1991; FERNANDES et al., 2011).

Additionally, the formation of reparative (tertiary) dentin, produced by odontoblasts as a defense mechanism against caries progression, further complicates age assessment. This process can appear radiographically similar to normal dental aging, leading to potential confusion. Such similarities are especially problematic when rehabilitated teeth are included in the sample, as they can obscure accurate age determination and introduce additional subjectivity (FARGES et al., 2015; RICUCCI et al., 2014).

Recent research has shown that Artificial Intelligence (AI) technologies, particularly Deep

Neural Networks (DNNs), present a promising approach to overcoming these limitations. By utilizing AI models to analyze panoramic radiographs, it is possible to achieve faster, more objective age estimates, minimizing dependence on manual evaluations and enhancing accuracy, particularly for older individuals (KIM et al., 2021; SHEN et al., 2021; GALIBOURG et al., 2021; SANTOSH et al., 2022; ZABOROWICZ et al., 2022; TOBEL et al., 2017; ŠTEPANOVSKỲ et al., 2017; AVUÇLU; BAŞÇIFTÇI, 2018; BANAR et al., 2020; BOEDI et al., 2020; VILA-BLANCO et al., 2020; HOU et al., 2021).

This dissertation addresses a relevant problem in Forensic Sciences by integrating radiological techniques with advanced AI methodologies to create a non-invasive, efficient, and less examiner-dependent approach to age estimation. However, when applying age estimation methods, including those based on AI, it is crucial to consider the specific geographical, socio-nutritional, and hormonal characteristics of the population being studied in order to reduce individual variability in dental development (GALIBOURG et al., 2021). This regional perspective is vital for improving the effectiveness of age estimations using dental radiographs.

Thus, the primary focus of this dissertation is to assess the geographic limitations of AI-based age estimation models by expanding our analysis across multiple regions of Brazil. In this context, we trained a machine learning model on patient data from one region and evaluated its performance on a dataset of patients from a different region to assess its robustness and ability to generalize across different regional populations. Additionally, we develop new approaches to create more efficient models that accommodate diverse regional characteristics, aiming to mitigate geographic biases and enhance performance.

The central hypothesis guiding this work is that when trained with regional data and adapted to different populations, AI models can effectively mitigate geographic biases and improve the accuracy of age estimation in various contexts. Our objectives include:

- Expanding the generalization capability of AI models for populations from different regions of Brazil.
- Establishing a multi-regional benchmark with Brazilian data to foster the development of robust models capable of handling diverse regional characteristics.
- Improving the adaptability of models to different regional characteristics by mitigating geographic biases.

This dissertation is organized as follows: Chapter 2 reviews the related work, providing an overview of the primary AI-based methods for age estimation, highlighting their contributions to different populations, and examining the challenges in achieving generalizable models. Chapter 3 presents the proposed approach of this study, detailing the motivations and improvements targeted by the research. Chapter 4 describes the methodologies used, including the data processing pipeline, the machine learning techniques, and the experiments conducted to validate the models' generalizability and robustness. It also presents a comprehensive analysis of each experiment conducted and observed biases. Chapter 5 discusses the results obtained from the experiments, analyzing the contributions and implications of the findings in the context of forensic age estimation. Finally, Chapter 6 concludes the study, summarizing the key findings, addressing the identified limitations, and proposing future research directions to further enhance the accuracy and generalizability of age estimation across diverse regional populations.

2 RELATED WORK

The current research aims to validate whether the methodologies and models developed using data from a single region can generalize effectively when applied to different regions within Brazil. By expanding the scope to encompass diverse Brazilian populations, this work assesses if the approaches developed for the northeast region of Brazil can maintain accuracy across the southeast region or if specific adaptations are required. If discrepancies are found, modifications to the model or the development of more generalizable neural network architectures will be investigated to address regional variations within a unified research framework.

This work is informed by a broader landscape of studies that explored age and biological sex estimation using orthopantomograms (OPGs). These works established critical benchmarks and highlighted challenges that remain to be addressed. A summary of these related experiments is presented in Table 1, showcasing the evolution of techniques that this project builds upon and evaluates in a Brazilian context.

Table 1 – Overview of Main Related Works for Age Estimation with AI-Based Solutions

Related Work	Nationality	# Samples	Age Range	ROI
(TOBEL et al., 2017)	Belgian	400	7 - 24	3rd Molars
(ŠTEPANOVSKỲ et al., 2017)	Czech	976	2.7 - 20.2	Full PR
(AVUÇLU; BAŞÇIFTÇI, 2018)	Turkish	1313	4 - 63	All Molars
(BANAR et al., 2020)	Belgian	400	7 - 24	3rd Molars
(VILA-BLANCO et al., 2020)	Spanish	2289	4.5 - 89.2	Full PR
(KIM et al., 2021)	Korean	1586	0 - 60	All Molars
(SHEN et al., 2021)	Chinese	748	5 - 13	41-46
(HOU et al., 2021)	Chinese	27957	1 - 93	Full PR
(SANTOSH et al., 2022)	Indian	1142	1 - 70	Full PR

Source: The Author (2024). The table provides an overview of key studies on age estimation using artificial intelligence, highlighting the dataset's nationality, number of samples, age range of participants, and the region of interest (ROI) analyzed in each research.

(TOBEL et al., 2017) conducted a study on the mineralization stages of the lower left third molar to estimate age in a Belgian population. They used 400 panoramic radiographs of individuals aged 7 to 24, the research applied a modified Demirjian classification system to label 10 developmental stages. The preprocessing involved manual region-of-interest adjustments and bounding box creation via Photoshop. Transfer learning with a fine-tuned AlexNet model, initialized with ImageNet weights, achieved the best results among tested approaches. The

method demonstrated an accuracy of 51%, a mean absolute difference of 0.6 stages, and a weighted Kappa of 0.82. While promising, the study highlighted the need for further refinements, such as automating preprocessing, increasing dataset size, and extending the approach to include direct age estimation. This work marked a significant step toward reliable, automated dental age estimation methods in forensic odontology.

(ŠTEPANOVSKÝ et al., 2017) extended the exploration of age estimation methods by analyzing the mineralization stages of multiple teeth in a Czech population of 976 individuals aged 2.7 to 20.5 years. Using a comparative framework, they evaluated 22 methods, ranging from simple regression to advanced data mining approaches, highlighting the balance between accuracy and usability. Their findings emphasized that tabular multiple linear regression achieved strong performance (MAE $<$ 0.7 years) while remaining user-friendly, unlike more computationally demanding methods such as neural networks. Notably, Štepanovský's work introduced a KDD-style approach, integrating data from multiple teeth to enhance age estimation, and underscored the importance of selecting practical, interpretable models for forensic applications. These insights set the stage for further method refinement, particularly in handling missing data and generalizing across diverse populations.

(AVUÇLU; BAŞÇIFTÇI, 2018) proposed a novel approach leveraging image processing and machine learning for simultaneous age and gender estimation in a Turkish population. The authors used 1,313 panoramic radiographs spanning an age range of 4 to 63 years. Their study applied rigorous preprocessing steps, including deskewing, gray filtering, and segmentation into nine quadrants to isolate features. These features were then processed through a multilayer perceptron (MLP) neural network, achieving classification accuracies exceeding 99%. Their method dynamically adjusted inputs to optimize performance and demonstrated the effectiveness of segment-based feature extraction for robust predictions. This study emphasizes the potential for automated forensic age and gender estimation while setting a benchmark for integrating preprocessing techniques with machine learning frameworks.

Building on prior works, (TOBEL et al., 2017), (BANAR et al., 2020) further advanced the automation of third molar staging with a fully automated three-step deep learning workflow. The authors used a dataset of 400 panoramic radiographs from Belgian individuals aged 7 to 24. This study employed pre-trained AlexNet and DenseNet201 architectures for molar localization, segmentation, and classification. With preprocessing steps such as rotation-based augmentation, the proposed pipeline demonstrated high accuracy ($>$99%) and competitive MAE values, highlighting the efficiency and speed of automated staging compared to manual

efforts. However, the reliance on a limited, pre-selected dataset may limit generalizability, a challenge acknowledged by the authors for future exploration. This study underscores the growing potential of deep learning for forensic applications, with further refinements required to address variability and expand usability across diverse populations.

Extending the exploration of automated age estimation, (VILA-BLANCO et al., 2020) introduced a multi-task deep learning framework using two novel architectures, DANet and DASNet, for chronological age prediction from OPG images. The authors used a dataset of 2,289 Spanish individuals aged 4.5 to 89.2 years. Their approach incorporated sex-specific features to improve age prediction accuracy. DASNet, in particular, demonstrated superior performance across all metrics, achieving a mean absolute error of 2.84 $\pm$ 3.75 years and classification accuracy of 85.4% for sex prediction. The study emphasized the robustness of DASNet in handling diverse radiological qualities and conditioning dental characteristics while leveraging Grad-CAM to visualize regions that contribute most to inference. This work highlighted the advantages of integrating multi-task learning with attention mechanisms, setting a new benchmark in forensic age estimation research.

(KIM et al., 2021) introduced a CNN-based system for age-group classification using panoramic radiographs of the first molars (#16, #26, #36, and #46) from 1,586 Korean individuals aged 0 to 60 years. Their model, trained using ImageNet-based transfer learning, achieved classification accuracies between 94% and 98% across three and five age groupings, with Grad-CAM visualizations confirming the focus on relevant anatomical features such as pulp size, alveolar bone levels, and interdental spaces. By integrating augmentation and majority voting techniques, the system improved patient-wise predictions, emphasizing the robustness of ensemble learning. This work demonstrated significant advancements in applying AI to forensic age estimation, leveraging CNNs for both automated feature extraction and interpretable decision-making.

(SHEN et al., 2021) explored the application of machine learning models for dental age estimation in a dataset of 748 panoramic radiographs from Chinese children aged 5 to 13. Building on the Cameriere method, which uses apical measurements of the seven lower left permanent teeth, the study trained random forest, support vector machine, and linear regression models. The ML models outperformed the traditional Cameriere formula in all metrics, achieving mean errors close to zero and a mean absolute error of 0.489 years for the SVM model. By leveraging manually extracted features such as apex distance and tooth length, the models demonstrated significant improvements in prediction accuracy while maintaining interpretability. Although

the sample range was relatively narrow, the study highlighted the potential of integrating ML algorithms with established forensic methodologies to enhance precision and reliability.

(HOU et al., 2021) significantly contributed to age estimation research by introducing a dataset of 27,957 panoramic radiographs from Chinese individuals aged 1 to 93, with extensive age coverage and high label accuracy. The authors used advanced neural architecture search techniques. The study compared various deep neural network architectures, including ResNet, DenseNet, and NASNet, with NASNet achieving the best performance (MAE of 1.64 years). Key innovations included optimizing model depth, employing multi-branch architectures, and exploring convolutional kernel asymmetry to tailor networks to the unique characteristics of dental radiographs. The insights gained from this work, particularly regarding lightweight models and kernel design, demonstrated the value of NAS-based approaches in forensic dentistry.

(SANTOSH et al., 2022) proposed a machine learning-based framework for age and gender determination using a dataset of 1,142 panoramic radiographs from Indian individuals aged 1 to 70+ years. The study employed a preprocessing pipeline that included Gaussian filtering for noise reduction and edge detection for manual feature extraction, such as inter canine distance and incisor width. Multiclass SVM was utilized for age estimation across 11 age groups, while LIBSVM was employed for gender classification, achieving accuracies of 97% and 95%, respectively. The methodology emphasized the robustness of SVM models in handling multidimensional odontometric features while maintaining simplicity and interpretability. This work underscores the importance of leveraging feature engineering alongside classical machine learning for reliable forensic identification.

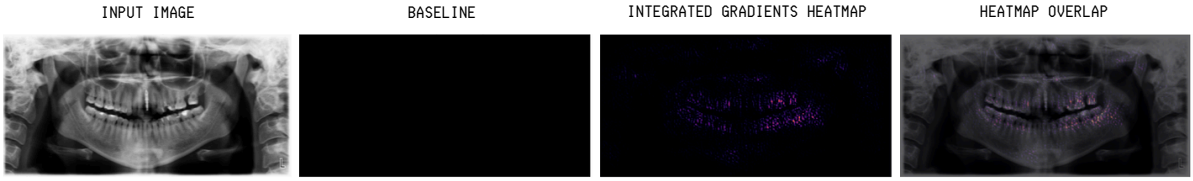
Research in computer vision has been pivotal in constructing the robust pipeline utilized in this study for training and validating machine learning models. An essential foundation was the adoption of Kaiming He initialization (HE et al., 2015), which ensures stable convergence in deep networks by addressing challenges associated with rectifier activations. This method was critical for training models efficiently and reliably on diverse dental datasets. Additionally, the use of combined asymmetric kernels, as introduced in the Inception-v4 architecture (SZEGEDY et al., 2017), significantly improved feature extraction by enabling efficient multi-scale analysis. This design closely aligns with the multi-kernel approaches explored by (HOU et al., 2021), allowing for a pipeline optimized for complex and heterogeneous radiographic features. Together, these advancements ensured scalability, stability, and high performance, which are critical for addressing the challenges inherent to dental image analysis.

Another critical contribution is exploring attribution techniques, as detailed by (SUN-

DARARAJAN; TALY; YAN, 2017). The Integrated Gradients (IG) method quantifies the contribution of each input feature to a model's prediction by calculating the integral of gradients along a linear path from a baseline input to the actual input. The baseline, often a neutral input like a completely black image, represents the absence of features. IG interpolates between the baseline and the actual input in small steps, computing gradients at each step, and accumulates these gradients to produce an importance score for each feature. This approach ensures that the attributions reflect the sensitivity of the model to changes in the input features. Guided by axiomatic principles such as Sensitivity, which ensures non-zero attribution for impactful features, and Implementation Invariance, which guarantees consistent attributions for functionally equivalent models, Integrated Gradients provides a principled and theoretically rigorous approach to interpreting model predictions.

A visualization of this process, such as heatmaps generated by IG, highlights the most influential regions of the input that contributed to the model's decision. This aligns with the use of Grad-CAM in dental radiography models, where attention maps validate the model's focus on clinically relevant regions and provide insights into its decision-making processes. For instance, in dental radiographs, IG heatmaps can emphasize regions like the first molars or jawbone areas critical for age estimation, enabling both interpretability and trustworthiness in model predictions. Figure 1 illustrates an example of Integrated Gradients applied to a dental radiograph, showing how specific regions are highlighted based on their contribution to the model's prediction.

Figure 1 – Integrated Gradients Exemple



Source: The Author (2024). Example of Integrated Gradients applied to a dental radiograph to the task of Age Estimation. The highlighted regions indicate features that contributed most to the model's prediction.

These attribution techniques bridge the gap between model performance and interpretability, enhancing their applicability in real-world medical contexts.

Additionally, the comprehensive survey by (MUMUNI; MUMUNI, 2022) on modern data augmentation techniques underscores the pivotal role of synthetic data and advanced transformations in addressing the limitations of small or imbalanced datasets. Strategies such as geometric

transformations and neural style transfer, highlighted in the survey, contribute to increasing diversity in training data, reducing overfitting, and improving model generalization. These augmentation methods are particularly relevant for dental datasets, where sample diversity often poses a challenge. By incorporating such augmentation strategies into the training pipeline, this study ensures a more robust model capable of generalizing across varied demographic and anatomical features.

The increasing interest in dental age assessment and the development of advanced methods combining dental radiology and neural networks has underscored the potential for localized studies to yield distinct results. Brazil's unique genetic diversity, shaped by centuries of migration from Asia, Africa, and Europe (MOURA et al., 2015), provides an opportunity to develop a well-balanced and representative sample, enhancing the generalizability of findings.

This study leverages this diversity to create a more balanced and inclusive benchmark compared to other studies. By utilizing a dataset that reflects such genetic variation, the research aims to validate the model's performance across different regions and establish a benchmark to assess models built for diverse populations.

3 PROPOSED APPROACH

This dissertation is part of a broader project involving the development of a dataset and models for age estimation using panoramic radiographs, which has been approved by the Ethics Committee of the Universidade Federal de Pernambuco and the Center for Medical Sciences under the Certificate of Ethical Appreciation Presentation (CAAE) N.º 42878921.6.0000.5208. The datasets used in this research include radiographs from two distinct sources: one from the Universidade Federal de Pernambuco (UFPE), representing the northeastern region, and another from the Faculdade de Odontologia de Piracicaba, Universidade de Campinas (UNICAMP), representing the southeastern region. These datasets are further described in Section 4.1, titled "Dataset Description and Distribution".

This study leverages these datasets to address the challenge of developing machine learning models capable of age estimation across geographically distinct populations. The methodology focuses on evaluating the limitations of models trained in one regional context when applied to data from another region and exploring strategies to enhance model generalization and robustness.

The proposed approach involves constructing a comprehensive experimental pipeline designed to systematically manage data processing, model training, and evaluation. This pipeline ensures that all experimental steps are consistent and reproducible, facilitating the validation of results. It integrates key components such as data augmentation, transfer learning, and region-aware model fine-tuning to address the identified challenges. Additionally, the pipeline supports experiments with training models from scratch on mixed datasets, aiming to balance predictive accuracy and generalization.

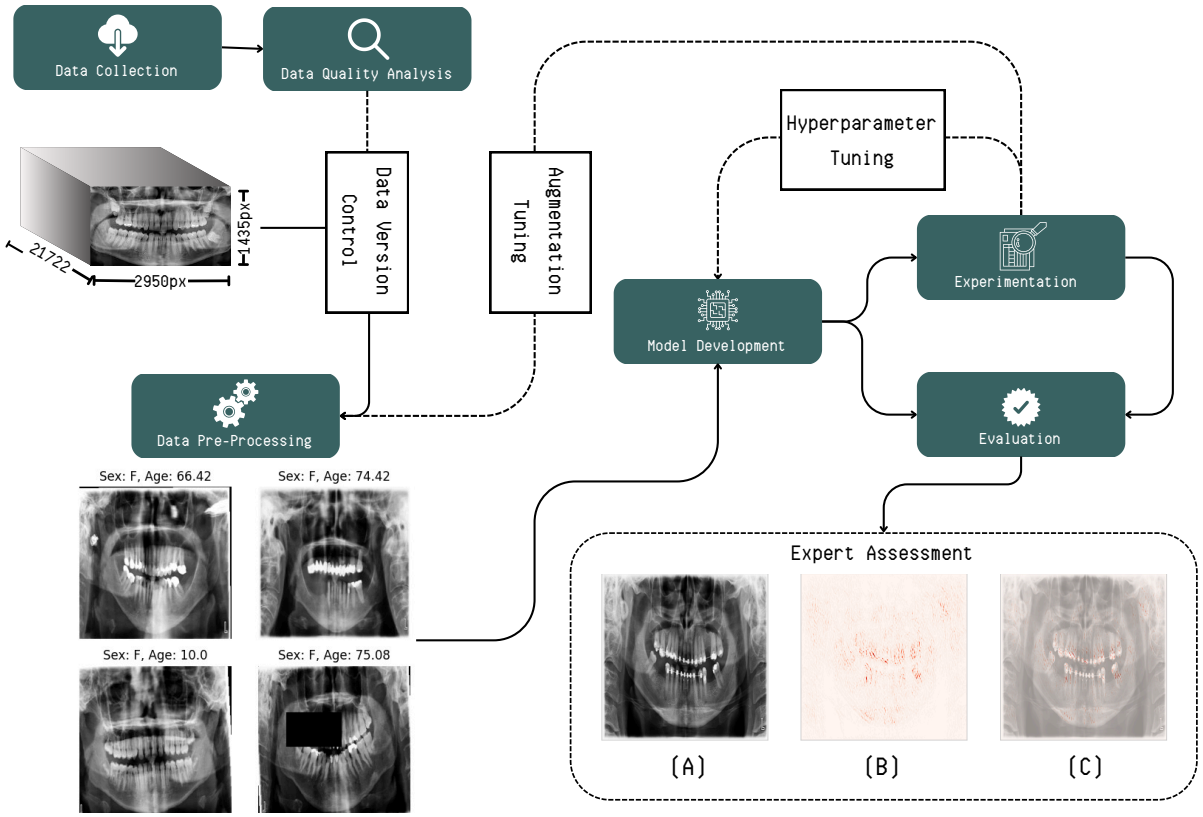
To ensure reliable input data for training, the radiographs underwent a quality review by experienced dentists specializing in radiology and forensic dentistry. Only images meeting clinical standards were included in the study.

The subsequent sections of this chapter describe the design and implementation of the experimental pipeline, detailing the specific methodologies and configurations employed in this research.

3.1 STUDY PIPELINE

The proposed pipeline, shown in Figure 2, encompasses several key stages, each meticulously designed to enhance the accuracy and robustness of our predictive models. These stages, outlined below, ensure a systematic approach, balancing the depth of data exploration and the precision of modeling techniques. This rigorous structure not only aids us in predicting chronological age but also sets the foundation for future project explorations, such as predicting biological sex.

Figure 2 – Proposed Pipeline



Source: The Author (2024). Schematic of the Data Processing and Model Development Pipeline. This pipeline demonstrates the six main steps of our process to perform the model training. In this example, the output of our solution is: (A) Original Image Age: 46.58 years; (B) Integrated Gradient Heatmap; (C) Model Predicted Age: 47.48 years

- Data Collection: Our research involved carefully collecting 21,722 panoramic radiograph images and corresponding patient information from the biobase of the UFPE and FOP/UNICAMP. A custom-built web scraping tool was employed to automate and streamline this data acquisition, ensuring both efficiency and the preservation of data

quality throughout the process. This meticulous approach was crucial for guaranteeing the reliability and validity of our dataset, which forms the foundation of the study.

- Data Quality Analysis: After acquiring the data, we conducted a thorough quality analysis to ensure that all files met the required criteria for this study. Since the data acquisition process was performed via web scraping, a manual evaluation was necessary to validate that all retrieved files were indeed panoramic radiographs and that the metadata was correctly paired with their respective users. During this process, we eliminated corrupted images, cropped radiographs (e.g., evaluations showing only one quadrant of the jaw), duplicate exams, and patients with multiple examinations. Duplicate exams were identified by comparing the hash values of each image, ensuring precise matching of identical files. For patients with multiple examinations, the most recent exam was retained based on the metadata provided by the biobank, which included information such as patient ID and examination date. This approach ensured consistency in the dataset and reduced potential biases arising from repeated examinations. Additionally, images were hashed to ensure data tracing, facilitating reproducibility and auditability throughout our experimental pipeline.

- DataLoader Structuring: We created custom DataLoaders to standardize the image sizes to 299x299 pixels, normalize pixel values to the range [0, 1], and apply data augmentation techniques on the training set. The DataLoaders were also configured to batch the images into sets of 32 for efficient feeding into our machine learning models. Further details about the preprocessing steps are provided in the Data Preprocessing Section.

- Model Construction: We trained multiple versions of a modified InceptionV4 model, leveraging state-of-the-art techniques from computer vision. These models were developed using automated pipelines incorporating best practices such as weight initialization, dynamic learning rate adjustments, and early stopping to mitigate overfitting.

- Structured Experimentation and Tunings: Structured experimentation was conducted using the Weights and Biases platform to optimize model performance. This included exploring various data augmentation strategies and fine-tuning hyperparameters within a controlled environment, allowing for systematic comparison and improvement of model versions.

- Model Evaluation: The pipeline's final stage involved evaluating our models' robustness and generalizability. This evaluation considered image degradation commonly present in panoramic radiographs and accounted for the specific characteristics of the Brazilian population. Collaboration with domain experts ensured that our models' performance was not only accurate but also interpretable and relevant across a wide variety of radiographs.

All experiments used Python 3.10.11 and PyTorch 2.0 to ensure consistency and reproducibility. To achieve deterministic results across experiments, persistent random seed settings (seed value of 0) were used. Additionally, all experiments were run on identical hardware—an Nvidia RTX 3060 GPU with 12GB VRAM—to maintain comparability.

Our experimental configuration involved several hyperparameters to fine-tune the model. We used an input size of 299x299 pixels for the InceptionV4 model, with a batch size of 32, and various data augmentation techniques, including horizontal flips, random brightness and contrast adjustments, rotations, translations, zoom, and pixel erasing. These transformations were applied with specific probabilities and factors to enhance the diversity of the training data. The model was trained for 100 epochs using the Adam optimizer with an initial learning rate of 0.001, adjusted via a ReduceLROnPlateau scheduler. Regularization methods such as Batch Normalization, Dropout (rate of 0.7), and Early Stopping (patience of 20 epochs) were employed to prevent overfitting. We used Mean Absolute Error (MAE) as the loss function, and the model was trained as a single task to predict age in years.

4 EXPERIMENTS

This chapter presents the dataset description, data pre-processing steps, model selection process, and a comprehensive overview of the experiments conducted to evaluate the geographic limitations of AI-based age estimation models. Each experiment was designed to progressively assess the model's adaptability to regional data variations and identify necessary adjustments to improve generalization across diverse Brazilian populations. The chapter is organized to first provide a detailed understanding of the data and methodologies, followed by the experimental setups and their results.

4.1 DATASET DESCRIPTION AND DISTRIBUTION

When developing methods for age estimation, particularly those based on Machine Learning, it is crucial to consider the target population's unique characteristics. These characteristics include geographical, socio-nutritional, and hormonal factors, significantly influencing dental development and introducing individual variability (GALIBOURG et al., 2021). Adopting a regional approach allows for a more precise and context-specific age estimation process using dental radiographs and is essential for reducing biases and enhancing model performance.

To further investigate the influence of regional differences on age estimation models, this study aims to evaluate whether models trained on data from the north-east region of Brazil, comprising 10,036 panoramic radiographs, can be effectively generalized to another dataset from Campinas, which consists of 11,686 panoramic radiographs collected under similar conditions. By analyzing the similarities and differences in model performance across these datasets, we seek to understand the degree of generalization achievable and identify potential adaptations or training strategies that could lead to more robust and generalizable models.

Both datasets include a diverse range of patients regarding age, sex, and dental development stages. Each image is labeled with a unique identifier ($image\_id$), and the XML file links this identifier to the corresponding age and sex information, verified using patients' identification records. This XML file provides a structured overview of the dataset's composition, enabling precise matching and easy retrieval of patient information during the training and validation procedures. To better illustrate the differences and similarities between the datasets from Unicamp and UFPE, we present a comparative analysis in Table 2:

Table 2 – Comparative Analysis of Datasets from Unicamp and UFPE with Age Statistics

	Category	Unicamp Dataset	UFPE Dataset
Total Exams	Total	11,686 (100%)	10,036 (100%)
	Male	5,020 (42.9%)	4,259 (42.4%)
	Female	6,666 (57.1%)	5,777 (57.6%)
Exams by Age Group	0-22 years	3,291 (28.2%)	2,505 (25.0%)
	22-65 years	7,364 (63.0%)	6,364 (63.4%)
	65+ years	1,031 (8.8%)	1,167 (11.6%)
Age Statistics	Mean Age	35.6 $\pm$ 20.0	38.2 $\pm$ 20.3
	Median Age	33.6	36.5
	IQR	32.4	32.3
	Minimum Age	2.1	2.25
	Maximum Age	89.1	96.5

Source: The Author (2024).

This table is instrumental in understanding the distribution of exams by sex and age group, highlighting key differences in representation between the datasets. Both datasets maintain a similar male-to-female ratio; however, there are notable discrepancies in age group representation, particularly at the extremes of the age range. Such insights are crucial for developing models that can generalize effectively across diverse populations and may suggest the need for specific model adjustments or weighting to address under-represented groups.

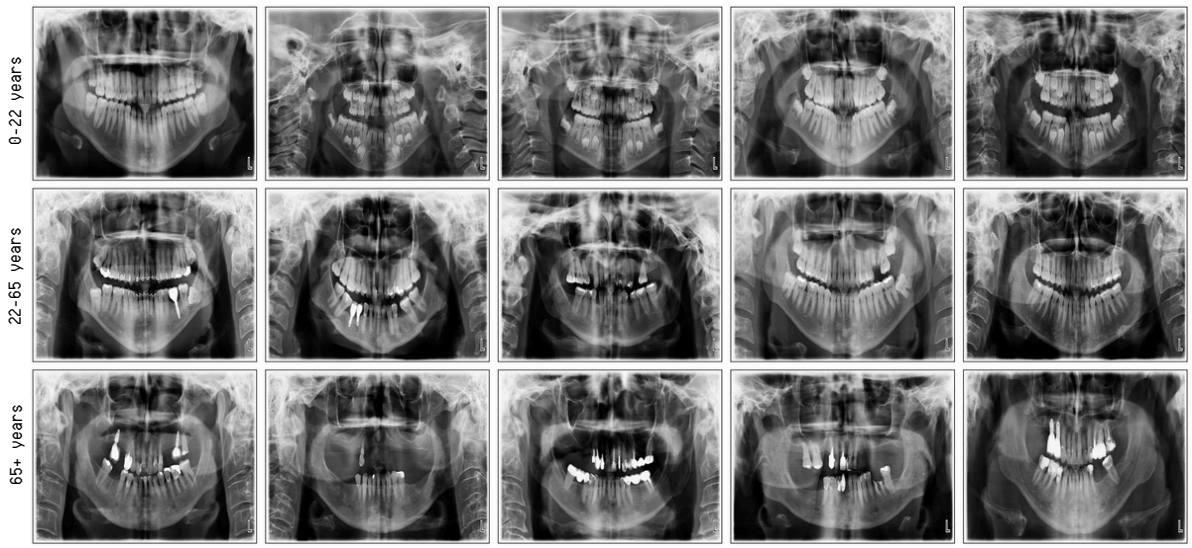
The comparative analysis of the datasets from Unicamp and UFPE provides valuable insights into their demographic structures and differences. The UFPE dataset has a slightly older population compared to Unicamp, with a mean age of 38.2 years versus 35.6 years and a median age of 36.5 years versus 33.6 years. This suggests that the UFPE dataset might include a relatively older cohort, which could impact the performance of age estimation models trained on this data. Additionally, the interquartile range (IQR) is very similar between the two datasets, indicating a comparable distribution for the middle 50% of ages. However, the maximum ages differ significantly, with the UFPE dataset ranging up to 96.5 years, while the Unicamp dataset has a maximum age of 89.1 years. The broader range in the UFPE dataset could pose challenges for model accuracy, particularly at the extreme age values due to increased variability.

Gender distribution is consistent across both datasets, with females comprising slightly more than half of the sample—57.1% in the Unicamp dataset and 57.6% in the UFPE dataset. This minor imbalance is unlikely to significantly affect model training, given the overall bal-

anced gender representation in both datasets. Age group analysis reveals that the majority of individuals in both datasets fall within the 22-65 age range, representing 63.0% in Unicamp and 63.4% in UFPE. Conversely, the 0-22 and 65+ age groups are underrepresented, which is an expected outcome due to the characteristics of dental clinics, where middle-aged patients are more likely to seek treatment, and this pattern aligns with the general population's healthcare-seeking behavior.

The high standard deviations (20.0 for Unicamp and 20.3 for UFPE) further underscore the considerable age variability within both datasets. This variability, combined with the underrepresentation of specific age groups, suggests that age-specific adjustments in model design could be beneficial to manage these challenges effectively. These demographic characteristics underscore the importance of tailored modeling approaches to ensure that age estimation models are accurate, equitable, and capable of generalizing across different age groups and both sexes.

Figure 3 – Sample from UFPE Dataset

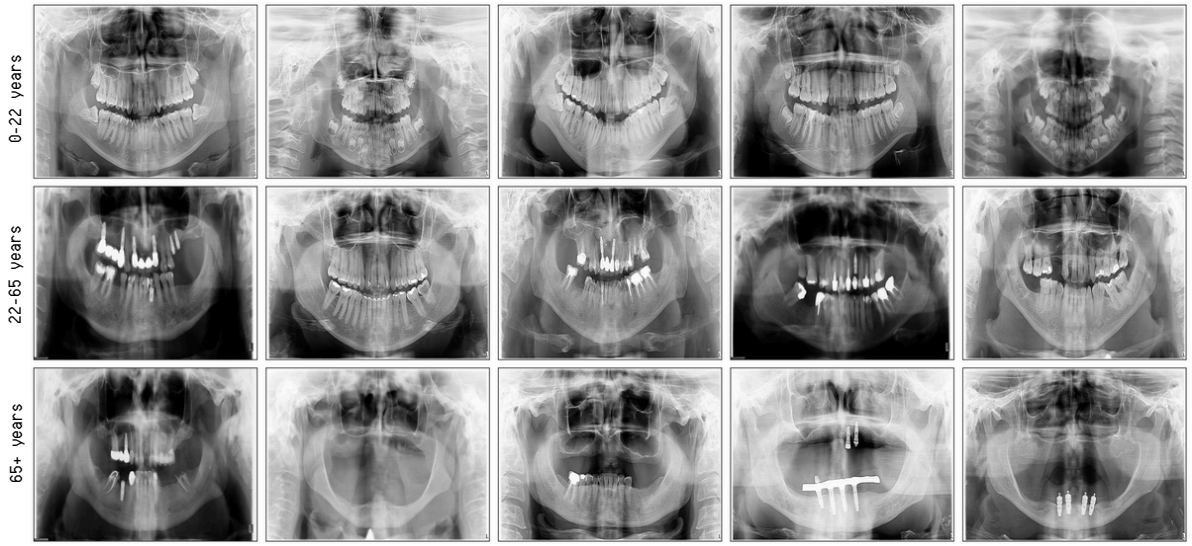


Source: The Author (2024).

The UFPE dataset (Figure 3) offers high-quality images with consistent contrast and clarity, making it particularly suitable for detailed analysis of bone structures. The images' sharpness and homogeneity clearly highlight dental and bony landmarks, which is crucial for AI models aiming to detect and evaluate these features accurately.

On the other hand, while also of high quality, the Unicamp dataset (Figure 4) presents images with softer contrast and a higher proportion of edentulous patients. This characteristic

Figure 4 – Sample from Unicamp Dataset



Source: The Author (2024).

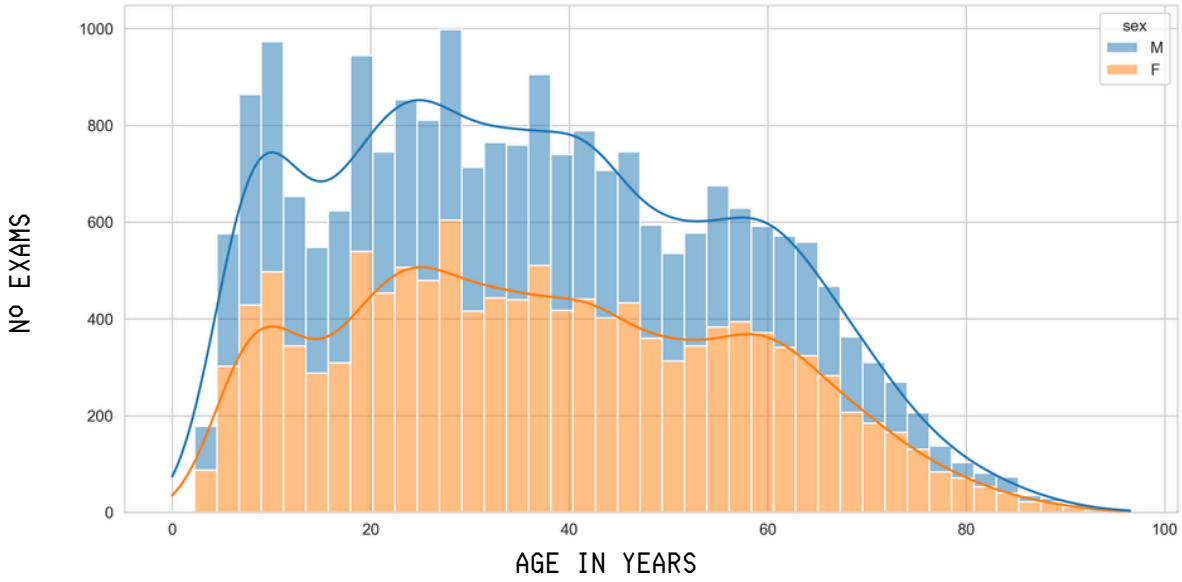
introduces an additional challenge for AI models, as many of the features commonly used for age estimation rely on the presence of teeth. In many cases, the absence of these dental landmarks may affect the model's ability to accurately interpret age-related bone structures. These differences in imaging parameters and patient populations suggest that AI models trained on one dataset might not achieve the same level of accuracy when applied to the other.

Additionally, the imaging equipment used in the datasets are different. The Unicamp dataset was acquired using the *OP100* and *OP300* X-ray machines from *Instrumentarium*, while the UFPE dataset utilized the *Planmeca ProMax*. These devices have distinct technical specifications, which likely contribute to the variations in image contrast and quality. Furthermore, the protocol settings and radiation exposure configurations in both datasets varied depending on the patient and the operator, making it impossible to establish a uniform imaging standard across the datasets.

The subtle differences between the UFPE and Unicamp datasets—such as variations in contrast, patient positioning, and the proportion of edentulous cases—could pose challenges for developing AI models that generalize well across both populations. These variations may impact the model's ability to consistently interpret bone structures and dental conditions in diverse imaging contexts.It is worth noting that the decision not to implement an automated image standardization process (e.g., contrast or brightness correction) was intentional. This approach aimed to encourage the model to adapt to varying imaging conditions, as standardization itself

would be equally complex and potentially non-adaptable to the diverse protocols and imaging setups encountered in new regions.

Figure 5 – Histogram of Patient Age Distribution by Sex Across Both Datasets



Source: The Author (2024).

To provide a comprehensive visual overview of the demographic distribution within the final dataset, which contains images from both UFPE and Unicamp, a histogram is presented to illustrate the age distribution by sex (see Figure 5). This visualization is particularly valuable for understanding how different age groups are represented and detecting potential biases or gaps in the dataset.

The histogram highlights the imbalance in the dataset, especially at the edges of the age spectrum—namely, the younger and older age groups. This imbalance is particularly pronounced in the older age range, where the steep decline in representation indicates a significant disparity.

Such distribution may hinder the model's ability to accurately predict these under-represented groups, especially older adults. Moreover, this imbalance suggests that the trained models might struggle to handle advanced-age cases effectively, even with adjustments like weighting strategies. Therefore, these insights reinforce the need to carefully consider model design to enhance robustness and ensure fair generalization across all age groups. If adjustments are not successful in addressing the challenges with advanced ages, it may be advisable to limit the model's usage to a specific age range where its performance is more reliable.

4.2 DATA PRE-PROCESSING

The preprocessing pipeline starts with creating a custom DataLoader, designed to standardize the dimensions of all images to 299x299 pixels and scale pixel values between 0 and 1. This normalization process is essential for ensuring that the input data is consistent, which is a critical factor in optimizing model performance and accuracy in image analysis.

The dataset is divided into 17,376 (80%) images for training, 2,173 (10%) for validation, and 2,173 (10%) as a holdout test set. The results presented in this study are always calculated based on this holdout test set, ensuring that the performance metrics reflect how the model performs on completely unseen data.

Extensive data augmentation techniques have been implemented to enhance the training dataset. These include image flipping, rotations, and modifications to brightness and contrast levels. By incorporating these transformations, it is possible to simulate a range of imaging conditions, which helps the model become more robust and capable of generalizing across different scenarios. More detailed information about the augmentation methods is provided in the subsection 4.2.1 on data augmentation. Pixel normalization is also maintained throughout the preprocessing pipeline to ensure uniformity across all images, which is vital for precise image processing.

The approach relies on Python scripts that manage these preprocessing steps using libraries such as Torchvision and PIL. These scripts handle image loading, applying augmentations, and saving the processed versions. Each image is resized to 299x299 pixels and normalized to align with the rigorous data processing requirements, ensuring consistent input quality for the models.

4.2.1 Augmentation Tuning

Drawing insights from the studies developed in (AVUÇLU; BAŞÇIFTÇI, 2018; BANAR et al., 2020; VILA-BLANCO et al., 2020; SANTOSH et al., 2022; MUMUNI; MUMUNI, 2022), and starting with our base model without augmentation as a foundation, we embarked on over 30 experimental trials. The objective was to discern the augmentation strategies that were most effective and could be seamlessly integrated into our DataLoader function. The configuration that resulted in the most promising outcomes is outlined in Table 3.

It's worth noting that the factor values for augmentation were intentionally configured to

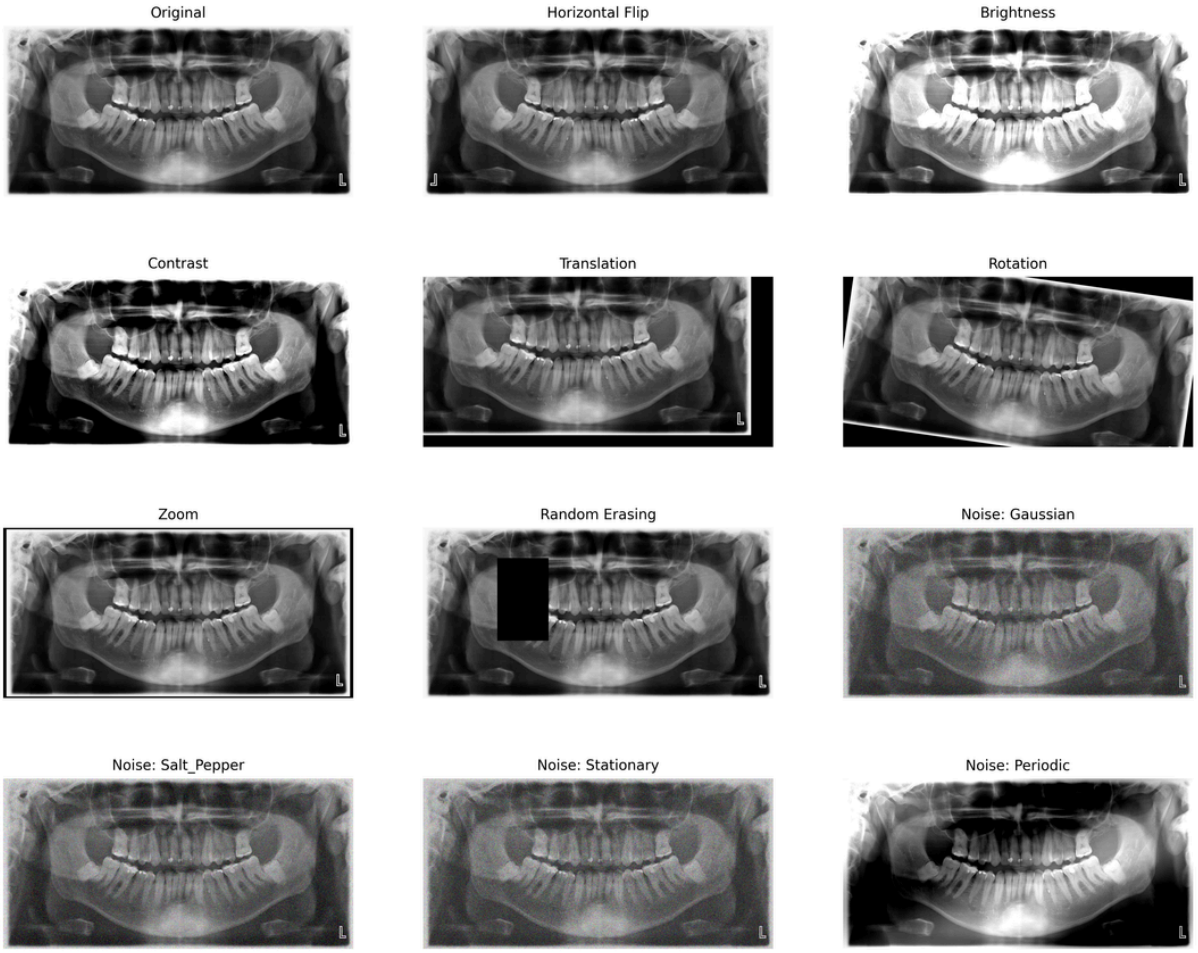
Table 3 – Data Augmentation Configuration

Transformation Name	**Probability**	**Factors**
Horizontal Flip	50%	-
Random Brightness	80%	0.15
Random Contrast	80%	0.15
Random Rotation	80%	3
Random Translation	80%	(0.1, 0.05)
Random Zoom	80%	(0.95, 1.05)
Random Erasing	15%	Scale: (0.05, 0.10) Ratio: (0.3, 3.3)
Random Noise:	50%	Intensity: (0.05, 0.2) Types: Gaussian Salt_Pepper Stationary Periodic

Source: The Author (2024).

effect minor alterations, aligning with findings in related literature, especially as recommended in (TOBEL et al., 2017; VILA-BLANCO et al., 2020). This decision was taken to strike a balance between introducing variability through augmentation and preserving the essential features inherent in the original images, given that obtaining OPGs is usually consistent and does not typically exhibit significant variability in patient positioning. These alterations can be observed in Figure 6, which demonstrates the impact of each augmentation strategy employed. It is important to highlight that a single image may undergo multiple transformations simultaneously, further enhancing the diversity within the dataset.

Figure 6 – Examples of Data Augmentation Strategies Applied to Panoramic Radiographs



Source: The Author (2024).

4.3 MODEL SELECTION

In (HOU et al., 2021), the researchers employed a Neural Architecture Search (NAS) approach to evaluate the impact of different architectural components, such as kernel configurations, multi-branch structures, architecture depth, and the utilization of pre-trained weights for transfer learning. Their findings showed that smaller architectures with multi-branch designs, asymmetric kernels, and no pre-trained weights generally perform better while reducing the risk of overfitting.

Various deep learning architectures are suitable for our age estimation tasks, including ResNet (HE et al., 2016), VGG (SIMONYAN; ZISSERMAN, 2014), and EfficientNet (TAN; LE, 2019).

ResNet, with its residual connections, enables the training of very deep networks by miti-

gating the vanishing gradient problem, which is crucial for learning complex features effectively. These connections act as shortcuts, allowing gradients to flow through the network more easily during training. However, this increased depth comes at the cost of computational resources, requiring substantial memory and time to train. Additionally, ResNet's depth can lead to overfitting for medical imaging datasets, which are typically smaller, making it less ideal without extensive data augmentation or regularization.

Following the analysis of ResNet, another architecture that has been considered for age estimation tasks is VGG. Although simpler, it does not utilize asymmetric kernels and follows a straightforward large architecture. This contrasts with the findings of Hou's NASNet, which demonstrated that smaller architectures with asymmetric kernels tend to perform better and reduce the risk of overfitting. For these reasons, we decided not to pursue the VGG architecture for our task.

EfficientNet uses a scaling approach that balances network depth, width, and resolution, providing a good trade-off between accuracy and computational efficiency. This makes EfficientNet suitable for many vision tasks, particularly when resources are constrained. However, it also lacks the multi-branch design that was found to be effective in Hou's study. Multi-branch architectures are better at capturing multi-scale features, which is especially important in age estimation from OPGs due to the mix of fine and coarse anatomical details present in the images. Thus, while EfficientNet offers strong performance overall, it may not provide the specialized feature extraction capabilities required for this task.

While modern architectures such as Swin Transformers (LIU et al., 2021) and Vision Transformers (ViT) (DOSOVITSKIY, 2020) have shown exceptional performance in various computer vision tasks, they were not selected for this study due to specific constraints related to the dataset size and computational resources. Transformers-based models typically require large-scale datasets to effectively learn the relationships between different image patches, which helps capture the global context. Moreover, the nature of OPGs often includes significant variability in image quality and anatomical features, making it challenging for transformer-based models to learn effectively without a substantial dataset size that provides diverse examples for generalization. Given the limited number of training images, applying such models would likely lead to overfitting, as they do not inherently possess the inductive biases present in Convolutional Neural Networks (CNNs), which make CNNs more efficient in generalizing from fewer samples. Additionally, the computational demands of training Transformer-based architectures are significantly higher, which makes them less practical for this particular study, where we are

resource-constrained.

In summary, while ResNet, VGG, EfficientNet, and Transformer-based models each offer unique strengths, their limitations—such as high computational demands, risk of overfitting with small datasets, lack of suitable kernel configurations, and architectural complexity—make them less ideal for our specific task. Instead, we chose the InceptionV4 (SZEGEDY et al., 2017) architecture for its effective balance between complexity, feature extraction capability, and computational efficiency. This architecture's multi-branch and asymmetric kernel design allows it to capture fine details and broader anatomical structures, which is crucial for analyzing OPGs and providing accurate age estimation. These capabilities help mitigate data scarcity and overfitting issues, aligning well with our study's requirements.

Guided by these findings, the initial approach for estimating chronological age from OPGs involved using the InceptionV4 architecture described in (SZEGEDY et al., 2017). The model was applied as an encoder without fine-tuning, followed by a dropout layer and two Fully Connected (FC) for feature decoding and the regression task. Despite its size, InceptionV4's use of multi-branch and asymmetric kernel configurations aligns with the recommendations from Hou et al.'s study, making it a suitable choice for this task by effectively balancing feature extraction and computational feasibility.

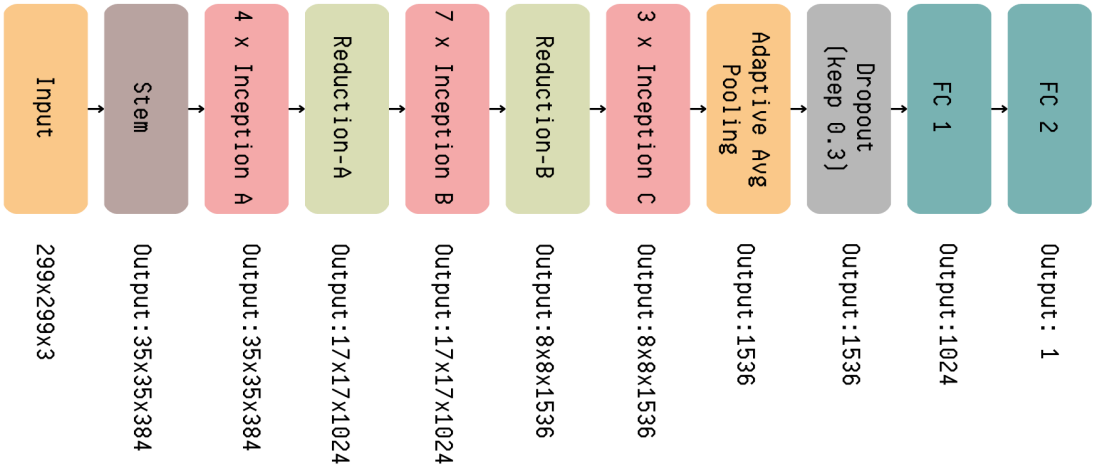
The detailed structure of the proposed network is depicted in Figure 7. The majority of layers employ the Rectified Linear Unit (ReLU) activation function, chosen for its ability to effectively mitigate the vanishing gradient problem, thus enabling efficient training (HU; ZHANG; GE, 2021). For the second FC layer responsible for the regression output, no activation function is used to allow for a direct linear transformation of inputs, ensuring a continuous output.

Figure 7 shows that the network starts with a 'Stem' module that performs initial convolutions to prepare the input image for multi-scale feature extraction. This module sets up a robust foundation for the subsequent stages of the architecture.

Following the 'Stem', the architecture incorporates the Inception-A module, which applies parallel convolutions of different types to capture a wide variety of features. This is succeeded by the Reduction-A module, which reduces the dimensions of the feature maps using pooling and stride-2 convolutions, enhancing computational efficiency.

The architecture then deploys the Inception-B module, which is similar in concept to Inception-A but uses different convolution configurations to capture more complex image features. This is followed by the Reduction-B module, which further reduces the dimensionality of the feature maps. Next, the Inception-C module refines these features, extracting more

Figure 7 – Schematic of the proposed network architecture



Source: The Author (2024). The Inception modules and their reduction mechanisms follow the structure outlined in (SZEGEDY et al., 2017). For more detailed information on the InceptionV4 components, please refer to the original publication.

detailed information crucial for the final task.

After this, an Adaptive Average Pooling layer condenses the spatial dimensions of the feature maps into a single vector, retaining key spatial information. A Dropout layer with a rate of 0.7 is applied post-pooling to prevent overfitting. By randomly deactivating some neurons during training, the model promotes a more balanced distribution of neuron weights.

The final part of the network includes two FC layers, FC1 and FC2, which decode the feature vector obtained from the preceding layers. FC1 interprets the features and identifies important patterns, while FC2 generates the final age prediction output based on the refined information from FC1.

To optimize the network's performance, the Kaiming Normal Method (HE et al., 2015), also known as He Initialization is used for weight initialization. This method is particularly effective with ReLU activation functions, ensuring that the variance of inputs remains consistent across layers and preventing issues such as vanishing or exploding gradients during backpropagation.

4.4 EXPERIMENTS OVERVIEW

Experiment 1: In this initial experiment, several models were trained using a regionally limited dataset (UFPE). The best-performing model was then tested on a distinct dataset (Unicamp) without any additional fine-tuning. This experiment establishes a baseline to assess the model's ability to generalize to geographically distinct populations without adaptations.

Experiment 2: Fine-Tuning with RMSprop on Limited Epochs Fine-tuning was conducted

using the RMSprop optimizer over a limited number of epochs to refine the model further. Two variations were tested: (1) fine-tuning the model using only the Unicamp dataset and (2) fine-tuning it with a combined dataset that includes images from both regions. This experiment aimed to examine whether minimal adjustments could improve performance in region-specific data or if a broader fine-tuning approach was more effective.

Experiment 3: Retraining on Combined Datasets In this experiment, the model was fully retrained on both datasets to examine its ability to generalize without relying on weights from the previous model. By retraining from scratch, this experiment provides insights into how a unified dataset affects the model's overall performance across different populations.

Experiment 4: Retraining with Augmentation on Combined Datasets The final experiment involved retraining the model on both datasets with data augmentation applied. Augmentation techniques were used to simulate various imaging conditions and further diversify the training data, enhancing the model's robustness. This experiment aimed to evaluate whether augmentation could improve the model's adaptability and performance on geographically diverse datasets.

Each of these experiments provides insights into how well the model performs across different regional data and identifies potential strategies for improving its generalizability.

4.5 **EXPERIMENT 1:** TESTING REGIONAL BIASES

4.5.1 Objective

This experiment aimed to establish a baseline by evaluating the performance of the best model trained on the UFPE dataset when applied to a new dataset from a different Brazilian region. The goal was to determine how well the model generalizes to geographically diverse data without any adaptations or fine-tuning.

4.5.2 Specific Methodology

For this initial test, the model was trained using the UFPE dataset, which comprised 10,036 panoramic radiographs representing a wide age range from 2.25 to 96.5 years. The training process involved the InceptionV4 architecture, chosen for its ability to effectively handle diverse anatomical structures using its multi-branch and asymmetric kernel design.

Throughout the experimentation phase, we executed over 30 variations of training configurations to optimize the model's performance. These variations included modifications to the data augmentation parameters, such as rotation angles, brightness levels, horizontal flips, as well as different augmentation multipliers. The objective was to determine the optimal configuration that would enhance the model's generalizability and robustness to variations in the imaging conditions present in panoramic radiographs.

The final model was trained for 100 epochs using the Adam optimizer with an initial learning rate of 10^{-3}, halving the learning rate at plateaus of 5 epochs to achieve gradual convergence. The Mean Absolute Error was used as the loss function, which is well-suited for continuous output like age estimation. Data augmentation was a crucial component of this process, incorporating random modifications to increase the diversity of the training dataset and reduce overfitting.

After extensive experimentation, the final model configuration was selected based on its superior performance in terms of accuracy and robustness on the validation set. This model was then applied directly to the new dataset from Unicamp without any further fine-tuning, allowing us to evaluate its generalizability to a geographically distinct population.

The general overview of the methodology for both training and testing phases is highlighted in Table 4:

Table 4 – Experiment 1 Methodology Overview - Training and Testing

	Component	**Description**
Training (UFPE Dataset)	Model Architecture	InceptionV4
	Dataset	UFPE
	Optimizer	Adam
	Learning Rate	10^{-3}
	Epochs	100
	Data Augmentation	Depicted in Table 3
Testing (Unicamp Dataset)	Dataset	Unicamp (Unseen data)
	Weights	Pre-trained from UFPE
	Optimizer	None
	Learning Rate	None
	Observation	No additional training applied

Source: The Author (2024).

4.5.3 Results

The model's performance on the new dataset, shown in Table 5, revealed a decline in accuracy compared to its performance on the original UFPE dataset. MAE increased to 4.97 years, while the MSE rose to 43.89 years^2, indicating reduced precision. Additionally, the R^2 score dropped to 0.888, with an explained variance of 0.915, reflecting the model's limitations in predicting age accurately for this new region. It is important to note that the metrics presented consider only the test subsets of both datasets, ensuring that the comparison is based solely on unseen data.

Table 5 – Experiment 1 results

Metric	**UFPE Dataset**	**Unicamp Dataset**
MAE	3.10 $\pm$ 0.18 years	4.97 $\pm$ 0.25 years
MSE	18.46 $\pm$ 0.27 years^2	43.89 $\pm$ 0.33 years^2
Median Absolute Error	2.16 years	3.88 years
IQR of Absolute Error	3.55 years	5.71 years
R^2	0.955	0.888
Explained Variance	0.965	0.915
T-statistic	1.88	-19.31
P-value	0.06	2.63×10^{-72}

Source: The Author (2024). Comparison of Performance Metrics for the best model trained on UFPE dataset versus Unicamp Dataset

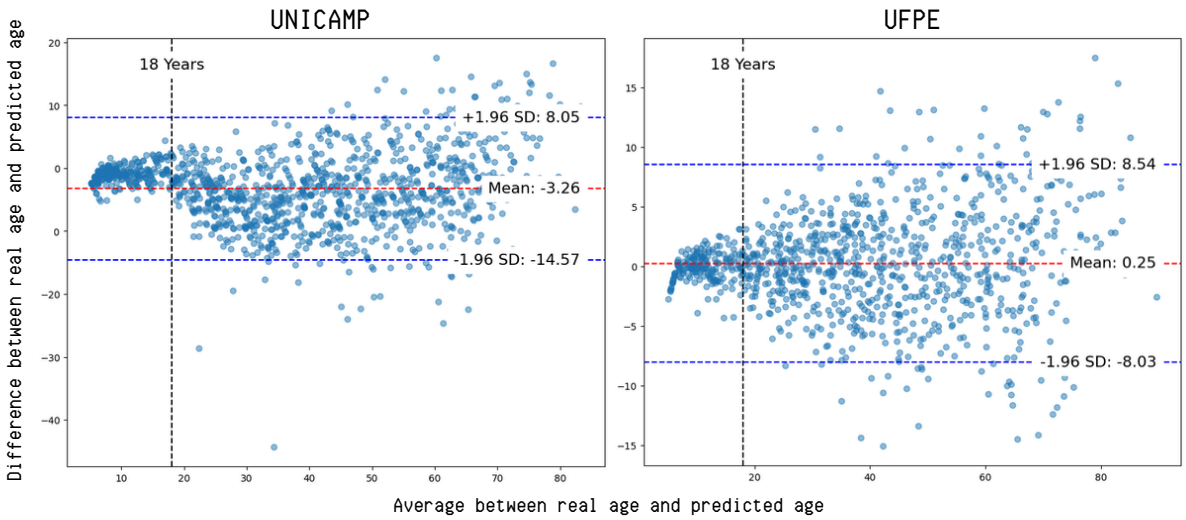
The Bland-Altman plot in Figure 8 visually demonstrates the agreement between real and predicted ages for both datasets.

For the Unicamp dataset, the mean difference of -3.26 years indicates a general tendency of the model to overestimate age. The limits of agreement, calculated at $\pm$ 1.96 standard deviations, span from -14.57 to 8.05 years, highlighting considerable variability in predictions. This variability is particularly pronounced in patients over 18 years, where deviations increase markedly.

While for the UFPE dataset we can observe a minor systematic bias in the prediction error distribution. This bias is indicated by the predominant clustering of values around 0.25 on the y-axis, and our confidence intervals exhibit a slight skew toward positive errors. This suggests that our model slightly underestimates the patients' age. However, based on the results from the t-test (p-value $= 0.06$), there is no solid statistical evidence to suggest that the predictions from our augmented model are significantly different from the actual ages.

The predictions on both datasets exhibited a cone-beam spread pattern, suggesting that the discrepancy between actual and predicted values widens as age increases. This observation implies that the model's predictive accuracy decreases with advancing age, likely due to the broad variability of odontological treatments, diseases, and other age-related changes. These factors contribute to increased complexity in accurately estimating age in older individuals.

Figure 8 – Bland-Altman Plot for Experiment 1



Source: The Author (2024).

A paired t-test was also conducted to evaluate the statistical significance of the differences between real and predicted ages. For the UFPE dataset, the resulting t-statistic and p-value indicated a value greater than 0.05, suggesting that there is no statistical evidence to support that the predicted ages are significantly different from the actual ages. In contrast, for the Unicamp dataset, the resulting t-statistic of -19.31 and p-value of 2.64×10^{-72} indicate a significant statistical discrepancy, confirming that the predicted ages differ considerably from the actual ages. With the results summarized above, we proceed to analyze the model's performance in detail, focusing on strengths, limitations, and specific failure cases.

4.5.4 Analysis and Discussion

The analysis of this baseline experiment highlights the model's strengths and limitations when applied to data from a new geographic region. Notably, the model tends to overestimate the ages of patients in the Unicamp dataset, a pattern supported by the Bland-Altman plot and further validated by the significant p-value from the t-test. This suggests that the predicted

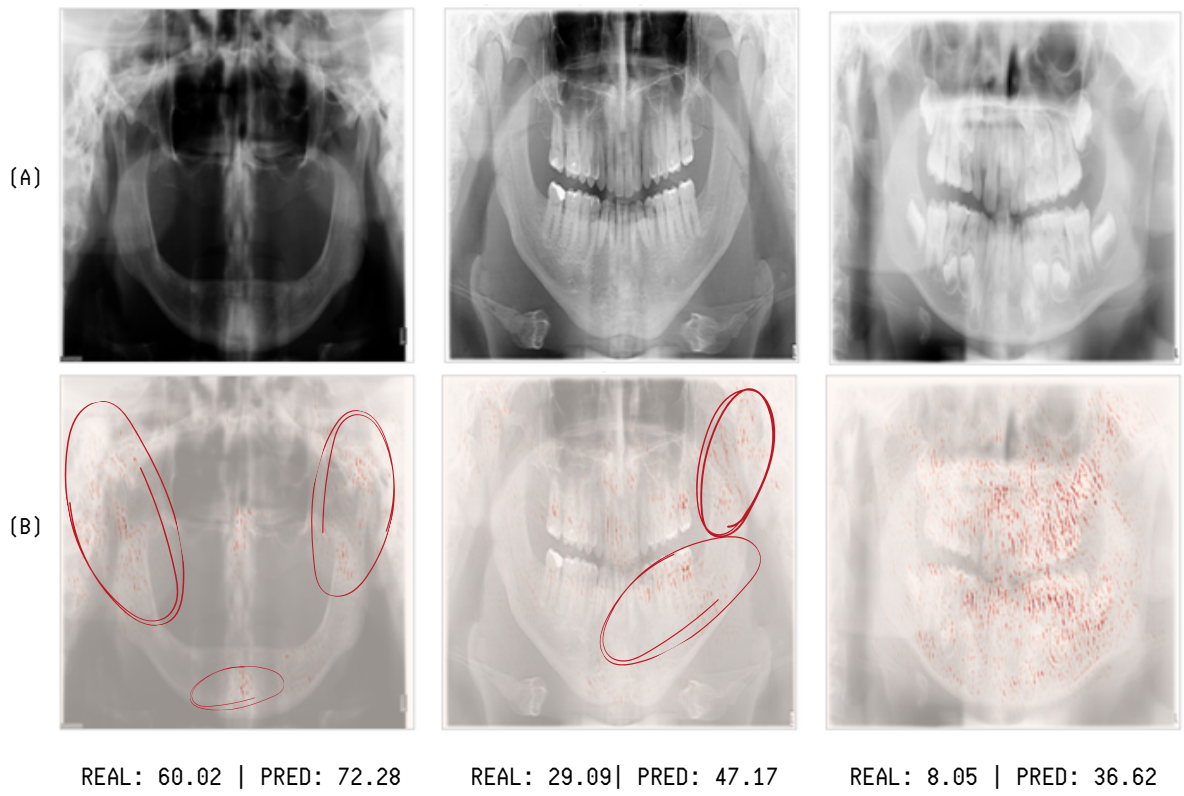
and actual ages do not share the same distribution, revealing a clear discrepancy in prediction accuracy.

When compared to the original model's performance on the UFPE dataset, prediction quality on the Unicamp data considerably declines. This reduction in accuracy indicates possible regional biases in the model's learned features, underscoring the need for further adaptation or fine-tuning to improve its generalizability across different Brazilian populations. Specifically, the significant drop in performance suggests that socio-environmental differences, such as nutrition and access to dental care, could impact dental development in ways not captured by the original training dataset.

Despite these challenges, the model demonstrated relatively strong performance in predicting ages for individuals up to 18 years, where growth patterns tend to be more consistent across regions. This indicates that some developmental features remain well-represented and generalizable, particularly in younger patients with lower variability.

To further understand the model's limitations, Figure 9 shows specific examples of failed predictions. These cases help illustrate typical scenarios where the model struggles:

Figure 9 – Examples of Model Failures in Age Prediction



Source: The Author (2024). Examples of cases where the model's predictions failed, highlighting common issues such as bone loss (Cases 1 and 2), and image overexposure (Case 3), shown from left to right.

- **Case 1**: This image represents a case of severe bone loss, likely due to early tooth loss or extraction, which exacerbated the condition, resulting in an overestimation of the patient's age.
- **Case 2**: In this case, the patient exhibits significant bone and tooth loss, which is not consistent with the typical characteristics expected for the patient's actual age, leading to a substantial prediction error.
- **Case 3**: The low exposure level of this image caused it to fall outside the distribution of images used during model training, even in the augmented dataset, which did not apply such severe augmentation in terms of brightness.

In conclusion, this experiment demonstrates that while the pre-trained model has specific strengths—particularly for younger patients—its limitations in terms of generalizability are evident when applied to data from new regions. Moving forward, steps such as fine-tuning the model with more regionally diverse data and employing advanced augmentation techniques will be crucial to improving its robustness and reducing geographic biases.

4.6 **EXPERIMENT 2:** FINE-TUNING PRE-TRAINED MODEL

4.6.1 Objective

This experiment aimed to explore the impact of fine-tuning the UFPE model in two different ways: (A) using only the Unicamp dataset and (B) using the combined UFPE and Unicamp datasets. The purpose was to determine if region-specific fine-tuning improves model performance for Unicamp without significant loss on the previous data and to assess whether including the UFPE dataset preserves generalizability across both datasets.

4.6.2 Specific Methodology

In this experiment, the model was fine-tuned over 20 epochs for each test, using both the Unicamp-only and combined datasets (UFPE and Unicamp). The goal was to evaluate the impact of different fine-tuning strategies on model performance across diverse data distributions. For both tests, we employed the InceptionV4 architecture with weights pre-trained from the

previous experiment to leverage the original model's learned features while adapting to new data.

The RMSprop optimizer was selected for fine-tuning, with a learning rate set at 3×10^{-5} to allow for gradual adjustments to the pre-trained model without risking abrupt changes. Additionally, a data augmentation strategy with a multiplier of 3 was applied to enhance the dataset's diversity, introducing brightness, contrast, and rotation variations, among other transformations. This augmentation was intended to improve model generalization across different patient demographics and imaging conditions. Table 6 summarizes the key components and parameters used in this experiment:

Table 6 – Experiment 2 Methodology Overview

Component	**Description**
Model Architecture	InceptionV4
Dataset	Both Datasets
Weights	Pre-trained from UFPE Experiments
Optimizer	RMSprop
Learning Rate	3×10^{-5}
Observation	Data Augmentation with a multiplier equal to 3

Source: The Author (2024).

4.6.3 Results

Table 7 summarizes the main performance metrics for both fine-tuning approaches applied across the UFPE and Unicamp datasets. When fine-tuned exclusively on the Unicamp dataset, the model achieved an MAE of 3.55 $\pm$ 0.20 years and an MSE of 24.13 $\pm$ 0.28 years^2 on Unicamp data, with an R^2 of 0.939. For the UFPE dataset under the same setup, the MAE increased to 3.61 $\pm$ 0.22 years, and the MSE rose to 25.36 $\pm$ 0.29 years^2, with an R^2 of 0.938.

In contrast, when fine-tuned on the combined dataset, the model achieved an MAE of 3.25 $\pm$ 0.20 years and an MSE of 20.47 $\pm$ 0.28 years^2 on the UFPE dataset, with an R^2 of 0.950. For the Unicamp dataset, the combined fine-tuning approach yielded an MAE of 3.69 $\pm$ 0.20 years and an MSE of 25.94 $\pm$ 0.29 years^2, with an R^2 of 0.934.

The Bland-Altman plots for the fine-tuned model using only Unicamp data (Figure 10) demonstrate the model's predictive accuracy and bias across the Unicamp and UFPE datasets. For the Unicamp dataset, the mean difference between real and predicted ages is close to zero

Table 7 – Experiment 2 Results

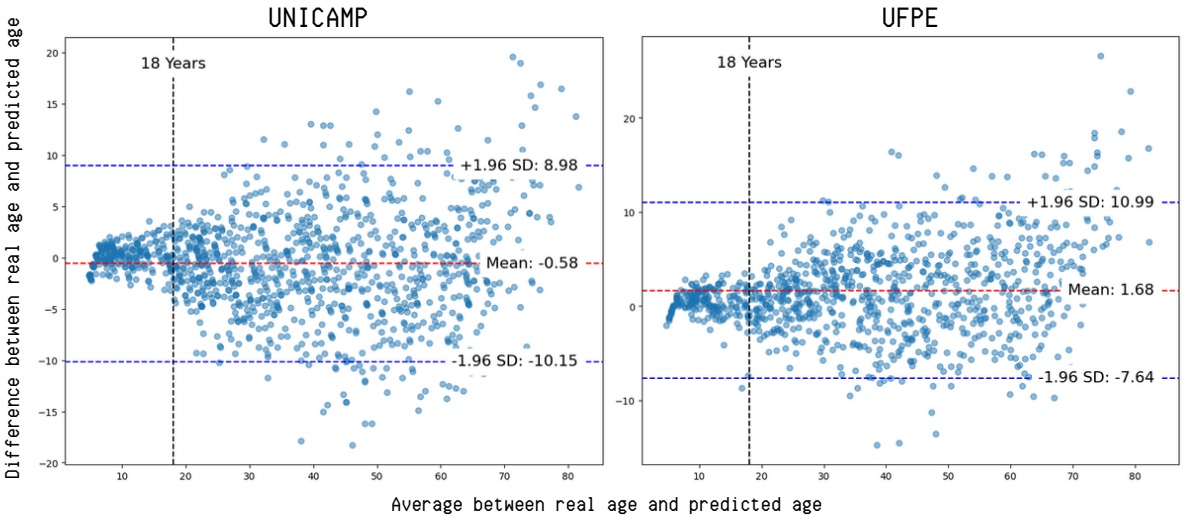
UFPE Dataset		
	Unicamp Only	**Full Dataset**
MAE	3.61 $\pm$ 0.22 years	3.25 $\pm$ 0.20 years
MSE	25.36 $\pm$ 0.29 years^2	20.47 $\pm$ 0.28 years^2
Median Absolute Error	2.51 years	2.21 years
IQR of Absolute Error	4.01 years	3.89 years
R^2	0.938	0.950
Explained Variance	0.945	0.950
T-statistic	11.18	0.34
P-value	2.00×10^{-27}	0.73
Unicamp Dataset		
	Unicamp Only	**Full Dataset**
MAE	3.55 $\pm$ 0.20 years	3.69 $\pm$ 0.20 years
MSE	24.13 $\pm$ 0.28 years^2	25.94 $\pm$ 0.29 years^2
Median Absolute Error	2.44 years	2.62 years
IQR of Absolute Error	4.09 years	4.36 years
R^2	0.939	0.934
Explained Variance	0.939	0.936
T-statistic	-4.09	-6.89
P-value	4.67×10^{-5}	9.16×10^{-12}

Source: The Author (2024). Results from Experiment 2 compare two fine-tuning strategies: one using only Unicamp data and the other using the combined training set (UFPE and Unicamp). The table highlights key performance metrics for each approach across the two datasets.

at -0.58 years, with limits of agreement ranging from -10.15 to 8.98 years, indicating tighter error dispersion around the mean. In contrast, for the UFPE dataset, the mean difference is slightly higher at 1.68 years, with limits of agreement from -7.64 to 10.99 years. This suggests a minor tendency to overestimate age for the UFPE dataset compared to the Unicamp dataset.

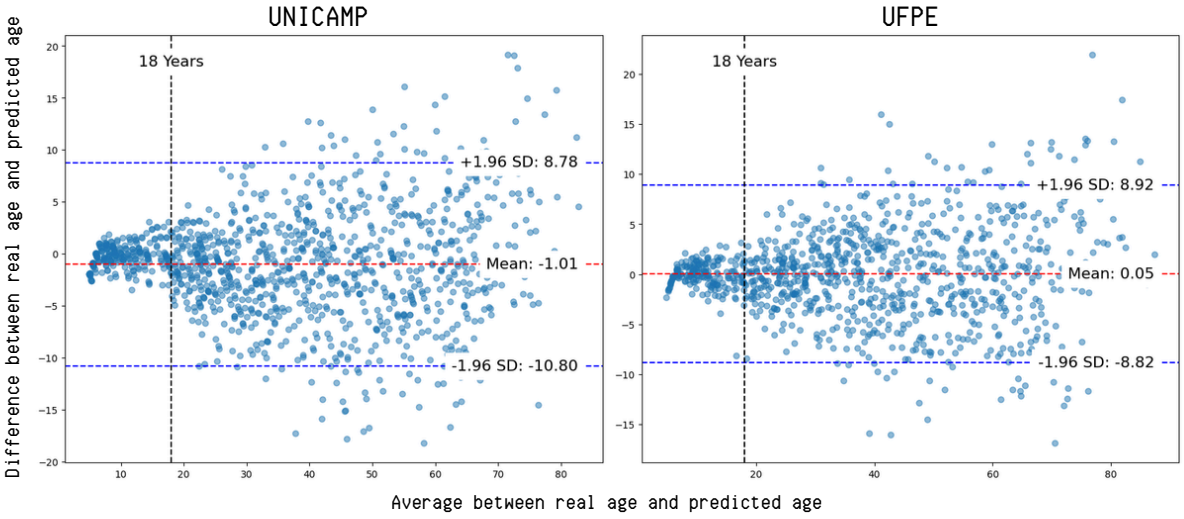
The Bland-Altman plots for the model fine-tuned using the complete training set (Figure 11) illustrate the model's predictive performance across both the Unicamp and UFPE datasets. For the Unicamp dataset, the mean difference between real and predicted ages is -1.01 years, with limits of agreement from -10.80 to 8.78 years, suggesting a slight underestimation bias. For the UFPE dataset, the mean difference is minimal at 0.05 years, with limits of agreement ranging from -8.82 to 8.92 years, indicating a balanced prediction with minimal bias.

Figure 10 – Bland-Altman Plot for Experiment 2 (A)



Source: The Author (2024). Comparison of results for the model fine-tuned with only Unicamp data across the Unicamp and UFPE datasets.

Figure 11 – Bland-Altman Plot for Experiment 2 (B)



Source: The Author (2024). Comparison of results for the model fine-tuned with the complete training set across the Unicamp and UFPE datasets.

4.6.4 Analysis and Discussion

The results from Experiment 2 show that fine-tuning the model, either with Unicamp-only data or the combined dataset, improved predictive accuracy compared to the baseline model used in Experiment 1. However, statistical testing using a Paired T-test indicates that, despite these improvements, the predicted ages are still statistically different from the actual ages, as evidenced by the p-values being significantly below 0.05 in most cases. This suggests that

while fine-tuning enhances model performance, it does not entirely eliminate the systematic differences in age prediction across regions.

When fine-tuning was performed exclusively on the Unicamp data, there was a marked improvement in the model's performance for the Unicamp dataset, with a reduction in both MAE and MSE. However, this region-specific tuning came at a cost: the model's accuracy declined considerably on the UFPE dataset, with increases in MAE and MSE and a noticeable deviation in the Bland-Altman plot. This indicates that the model has adapted to the Unicamp-specific characteristics, potentially losing its generalizability when applied to a different dataset like UFPE.

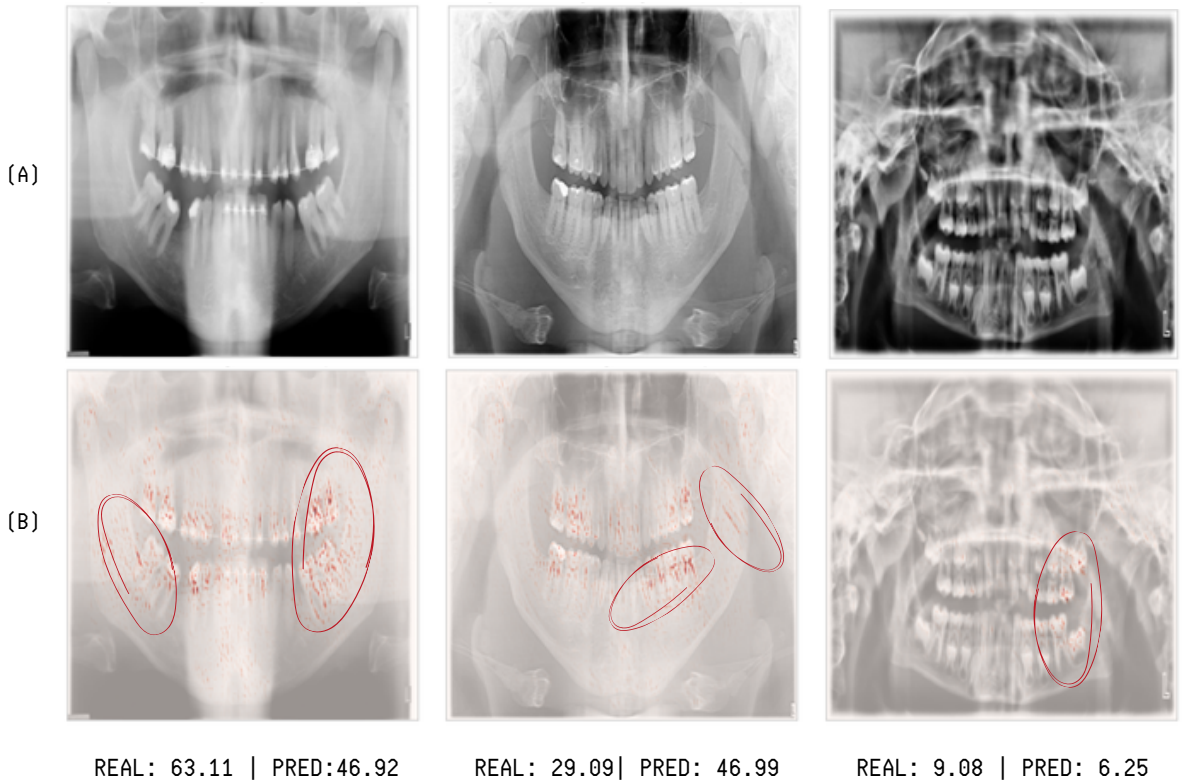
Conversely, fine-tuning on the combined dataset (both Unicamp and UFPE) achieved a more balanced performance across the two datasets. This approach yielded slightly higher MAE and MSE values for Unicamp than the Unicamp-only fine-tuning, but it maintained competitive performance on the UFPE dataset. The Bland-Altman plot for this combined approach showed a near-zero mean difference for the UFPE data, indicating minimal bias and a narrow range of limits of agreement, suggesting improved consistency across diverse data. However, the T-Test results still point to statistically significant differences between predicted and actual ages, underscoring persistent challenges in achieving true geographic generalizability.

The following figures show selected examples of prediction failures, along with their analysis by domain experts, to further illustrate the specific challenges the model encountered during Experiments 2 A and B.

Figures 12 and 13 show examples of failed predictions for both fine-tuning approaches, highlighting common scenarios where the model still struggles.

- Experiment 2 (A):
  - **Case 1**: The patient shows good overall dental condition, with minimal bone loss and wide dental pulp. This corresponds to a younger dental age than the actual chronological age, leading to an underestimation by the model.
  - **Case 2**: Similar to Experiment 1, the patient exhibits significant bone and tooth loss, which is not consistent with the typical characteristics expected for the patient's actual age, leading to a substantial prediction error.
  - **Case 3**: The image suggests that this acquisition was made using an adult protocol on a child. However, it is notable that the model focused on the correct region (first

Figure 12 – Examples of Model Failures in Age Prediction for Experiment 2 (A)

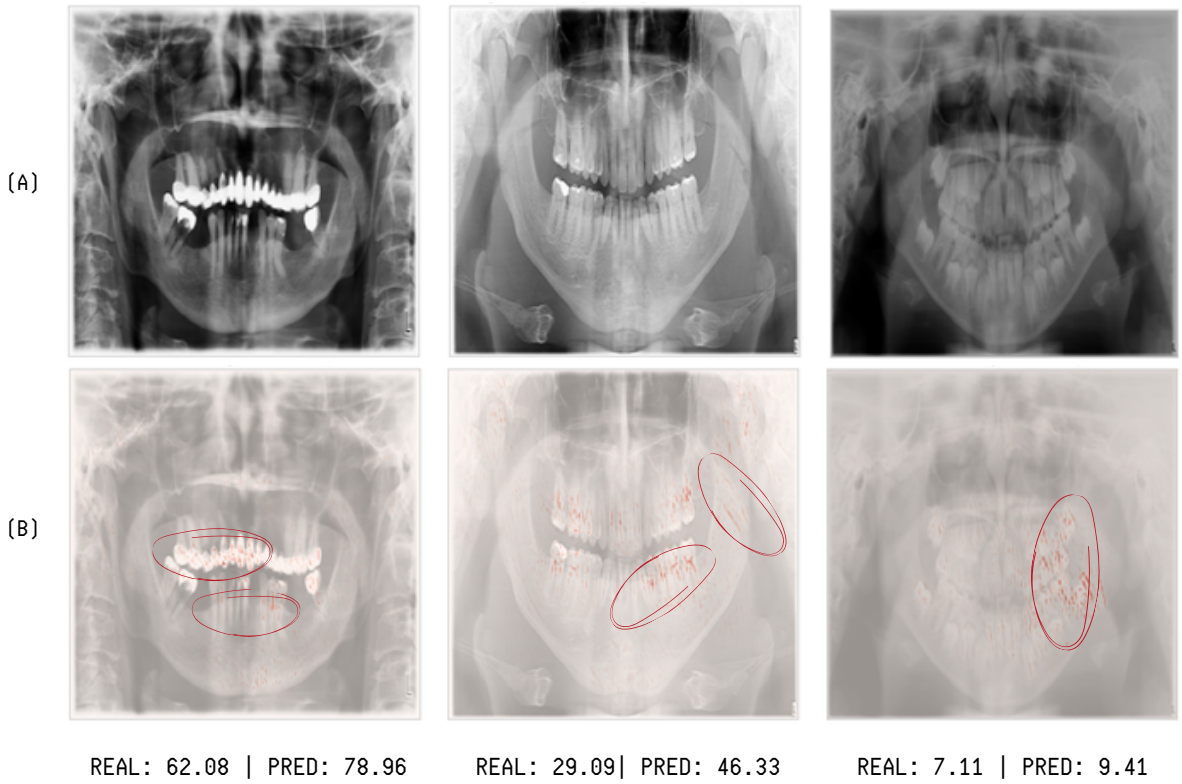


Source: The Author (2024). Examples of cases where the model's predictions failed in Experiments 2 (A), highlighting issues such as dental condition (Cases 1 and 2), and image quality (Case 3), shown from left to right.

molars), demonstrating its ability to identify relevant features despite the quality issues.

- Experiment 2 (B):
  - **Case 1**: The presence of crowns on all upper teeth indicates older dental treatments, which are no longer commonly used in younger patients. Additionally, highly calcified canals suggest a dental age older than the actual chronological age, leading to an overestimation by the model.
  - **Case 2**: Once again, the patient exhibits significant bone and tooth loss, which is not consistent with the typical characteristics expected for the patient's actual age, leading to a substantial prediction error.
  - **Case 3**: Despite the low density image, the model focused on the correct region, identifying the eruption of the first molar and the upward movement of the second molar, which are indicative of the 8-9 year age range.

Figure 13 – Examples of Model Failures in Age Prediction for Experiment 2 (B)



Source: The Author (2024). Examples of cases where the model's predictions failed in Experiments 2 (B), highlighting issues such as dental restoration procedures (Case 1), bone loss (Case 2), and image quality (Case 3), shown from left to right.

These failure cases provide insight into the specific situations where the model's predictions diverged from actual ages. Factors such as patient positioning, dental restoration procedures, and image quality significantly contributed to errors.

In summary, while both fine-tuning strategies improved model accuracy relative to the baseline, they also introduced trade-offs. Fine-tuning on Unicamp data alone enhances local accuracy but sacrifices performance on external datasets. The combined fine-tuning approach offers more balanced performance across datasets. However, statistically significant prediction errors remain, indicating the need for further adjustments or alternative approaches to achieve robust, geographically generalizable age estimation models.

4.7 **EXPERIMENT 3:** FULL RETRAINING ON BOTH DATASETS

4.7.1 Objective

The objective of Experiment 3 was to assess the impact of retraining the model from scratch using both the UFPE and Unicamp datasets. By eliminating any pre-trained weights and starting with a fresh model, this experiment aimed to evaluate whether a complete retraining approach could improve the model's ability to generalize across both datasets without inheriting any potential biases from prior training on a single region.

4.7.2 Specific Methodology

For this experiment, the model architecture remained the same-InceptionV4-, but it was initialized without any pre-trained weights. The dataset included both Unicamp and UFPE images to encourage balanced learning across different regions. Table 8 outlines the specific parameters used in this retraining approach:

Table 8 – Experiment 3 Methodology Overview

Component	**Description**
Model Architecture	InceptionV4
Dataset	Combined Dataset (UFPE and Unicamp)
Weights	Kaiming Initialization (HE et al., 2015)
Optimizer	Adam
Learning Rate	1×10^{-3}
Observation	Base model, no augmentation applied

Source: The Author (2024).

4.7.3 Results

Table 9 summarizes the main performance metrics for the retrained model on both the Unicamp and UFPE datasets. We achieved balanced performance across the two datasets by training the model from scratch on a combined dataset, suggesting that this approach effectively reduced regional biases.

For the Unicamp dataset, the retrained model obtained an MAE of 3.48 $\pm$ 0.19 years and

an MSE of 23.53 $\pm$ 0.28 years^2, with an R^2 score of 0.940 and an explained variance of 0.940. The median absolute error was 2.46 years, with an IQR of 4.10 years, indicating a consistent error distribution across age predictions.

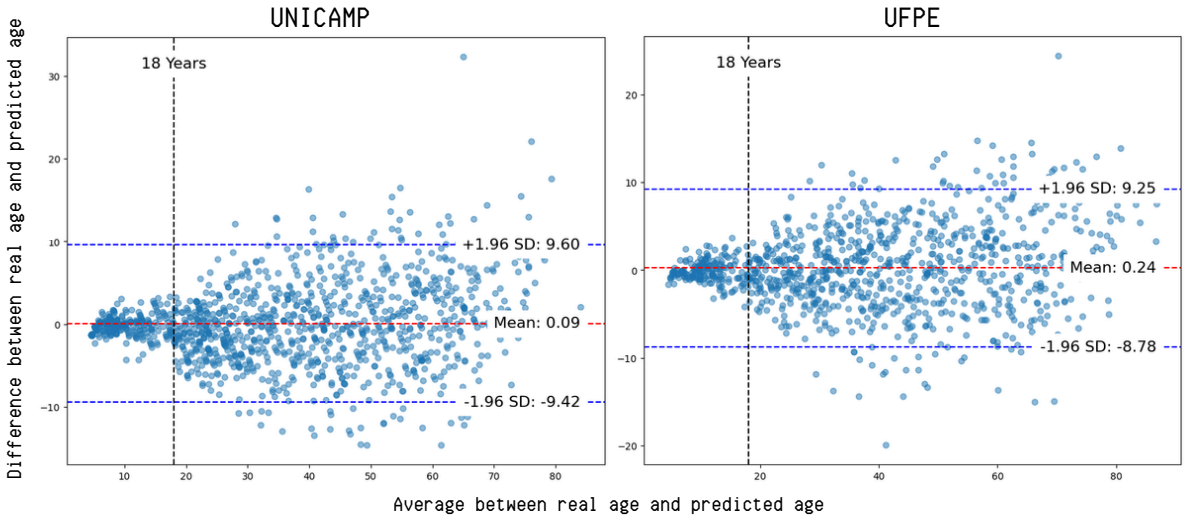
The model performed similarly on the UFPE dataset, achieving an MAE of 3.36 $\pm$ 0.20 years and an MSE of 21.19 $\pm$ 0.28 years^2, with an R^2 of 0.948 and an explained variance at 0.948. The median absolute error for this dataset was 2.48 years, with an IQR of 3.82 years, reflecting an equally stable error distribution.

Table 9 – Experiment 3 Results: Retraining Model on Combined Datasets)

Performance Metrics for Retrained Model		
	Unicamp Dataset	**UFPE Dataset**
MAE	3.48 $\pm$ 0.19 years	3.36 $\pm$ 0.20 years
MSE	23.53 $\pm$ 0.28 years^2	21.19 $\pm$ 0.28 years^2
Median Absolute Error	2.46 years	2.48 years
IQR of Absolute Error	4.10 years	3.82 years
R^2	0.940	0.948
Explained Variance	0.940	0.948
T-statistic	0.64	1.64
P-value	0.525	0.101

Source: The Author (2024). Summary of key performance metrics for Experiment 3, where the model was trained from scratch using a combined dataset.

Figure 14 – Bland-Altman Plot for Experiment 3



Source: The Author (2024). Comparison of results for the re-trained model across the Unicamp and UFPE datasets.

Additionally, the Bland-Altman plot, presented in Figure 14, shows a mean difference close

to zero for both datasets, with limits of agreement ranging from -9.42 to 9.60 years for Unicamp and -8.78 to 9.25 years for UFPE. This indicates minimal prediction bias, confirming that the model produces consistent age estimations across diverse regional data.

4.7.4 Analysis and Discussion

The results from Experiment 3 demonstrate that training the model from scratch on a combined dataset significantly enhanced its generalizability across both regions. The balanced performance on the Unicamp and UFPE datasets, reflected in the low MAE and high R^2 scores, suggests that incorporating data from both regions allowed the model to capture a broader representation of age-related features, reducing overfitting to any single population.

The Bland-Altman plots further support this improvement, showing minimal bias in predictions across both datasets. This outcome contrasts with Experiments 1 and 2, where models fine-tuned on individual datasets showed higher levels of regional bias. By training from scratch on a mixed dataset, the model appears to have successfully mitigated regional discrepancies, achieving reliable predictions across diverse Brazilian demographics.

Additionally, the T-Test results offer further insight into prediction alignment, with p-values (0.5251 for Unicamp and 0.1011 for UFPE) above the 0.05 threshold, suggesting no evidence to reject the null hypothesis. Unlike previous experiments, where fine-tuning on a single dataset led to statistically significant differences, this retrained model shows no detectable discrepancy between predicted and actual age distributions in either dataset.

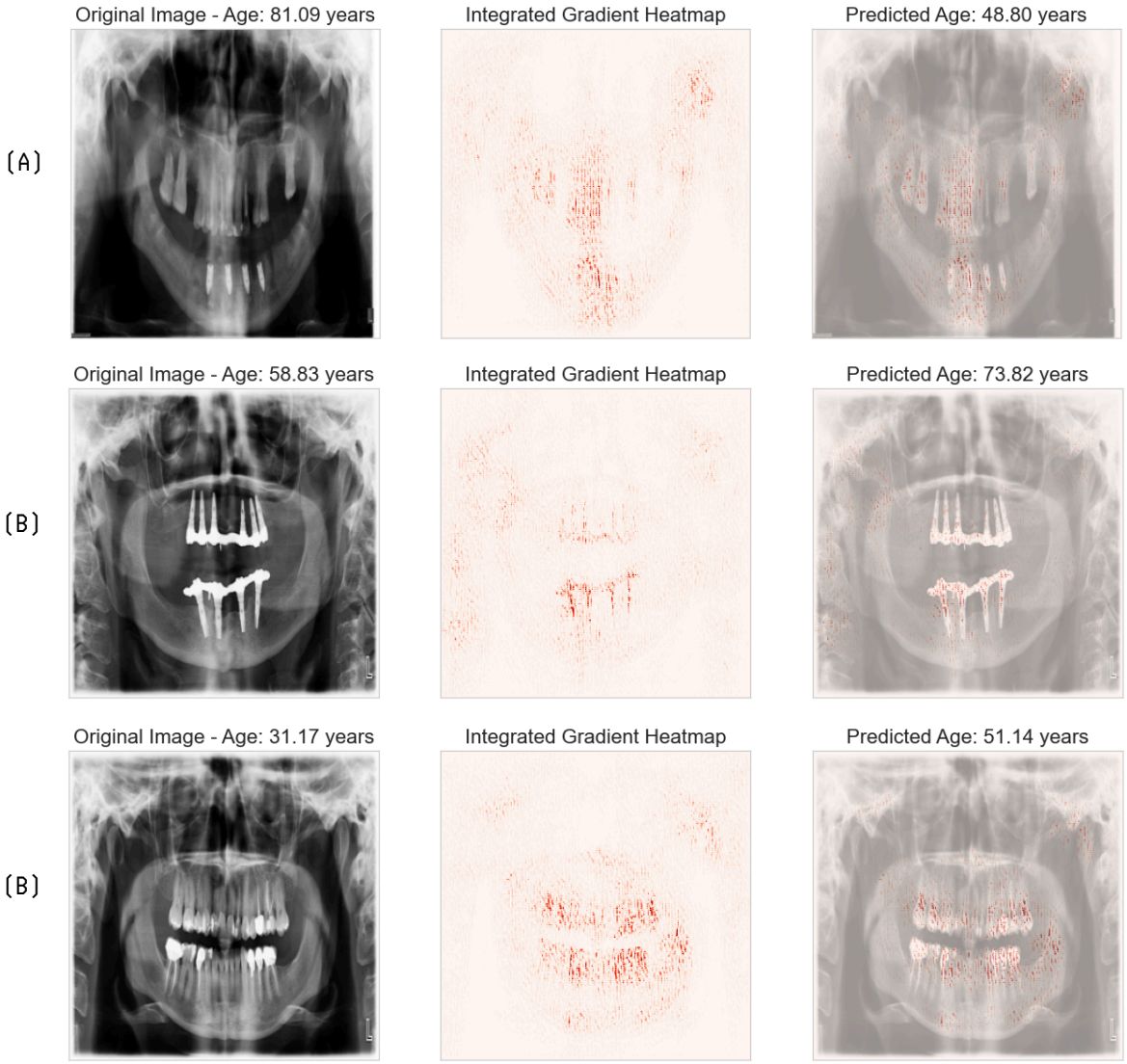
In summary, full retraining on the combined dataset resulted in a well-balanced model that generalizes effectively across both the UFPE and Unicamp datasets. This approach minimized regional bias, supporting the benefits of using a diverse dataset to enhance model robustness and consistency in age estimation across varied populations.

Significant outliers continue to appear despite these improvements, with some predictions deviating by over 20 years from the actual ages. While retraining reduced the frequency of these high-magnitude errors, they persist, particularly in cases involving elderly individuals, edentulous young patients, and instances where image quality issues affected the results.

These examples are further analyzed through the insights provided by a dental specialist, highlighting how specific artifacts and dental conditions impacted the model's predictions, as visually indicated by the integrated gradient heatmaps.

Figure 15 illustrates examples of these persistent outliers, highlighting how specific fac-

Figure 15 – Outlier Prediction Examples



Source: The Author (2024). Examples of prediction outliers highlight typical cases where the model's age estimations deviate significantly from the actual age. (A) Elderly patient with an actual age of 81.09 years but a predicted age of 48.80 years, illustrating the model's challenges with image artifacts, including a central white blur which obscures critical features. (B) Edentulous middle-aged patient, aged 58.83 years, predicted as 73.82 years, where the absence of teeth and dental markers contributes to the discrepancy. (C) Young adult, aged 31.17 years, with a predicted age of 51.14 years, demonstrating the influence of edentulous areas as a significant feature for the model in this specific case; this characteristic is atypical for young patients and may have misled the model's prediction. Each row displays the original image, the integrated gradient heatmap, and the overlaid prediction heatmap, which reveal regions the model focused on during prediction.

tors, such as extreme patient age, dental conditions, and image quality, contribute to notable prediction inaccuracies.

The insights from a dental specialist provided additional context regarding the specific factors influencing the errors in these outlier predictions:

For Case (A), the central artifact observed may be attributed to the lower implants, which affected the model's ability to extract relevant features accurately.

In Case (B), the use of double protocol prosthesis is typically seen in elderly patients, reinforcing why the model predicted a higher age for this middle-aged individual. The lack of dental markers, coupled with the presence of such prosthetics, likely contributed to the model's overestimation.

As for Case (C), the patient's oral condition appears older due to dental protection responses, such as the formation of tertiary dentin, possibly resulting from multiple cavities during childhood. This accumulation of calcified canals misled the model in predicting a significantly older age than the actual age, reflecting the challenge in distinguishing between dental evolution from age versus that from other factors like past dental caries.

These observations highlight the need for more sophisticated strategies, such as specific data augmentation targeting these conditions or the use of additional features to better capture such edge cases. These strategies would ensure that the model can reliably estimate age even in challenging scenarios.

Building on these findings, the following experiment was developed to investigate whether data augmentation techniques can further improve model robustness and reduce significant outliers. By introducing controlled variations in imaging conditions—such as brightness adjustments, noise injection, and random erasing—this experiment aims to expand the diversity of the training dataset.

4.8 **EXPERIMENT 4:** RETRAINING WITH DATA AUGMENTATION

4.8.1 Objective

The goal of Experiment 4 was to evaluate the impact of data augmentation on model performance, particularly in reducing high-magnitude prediction errors observed in Experiment 3. Given the persistent outliers, especially among extreme cases like elderly patients, young edentulous individuals, and images with artifacts, this experiment artificially introduced more variability in training data to enhance the model's resilience. We sought to diversify the training conditions by applying various augmentation techniques and improve the model's generalization capability across unusual patient profiles and imaging inconsistencies.

4.8.2 Specific Methodology

In this experiment, the InceptionV4 architecture was again trained from scratch using a combined dataset of UFPE and Unicamp images. Data augmentation techniques were applied extensively to increase image diversity, simulating various real-world imaging scenarios. Table 10 provides an overview of the model training parameters.

Table 10 – Experiment 4 Methodology Overview

Component	**Description**
Model Architecture	InceptionV4
Dataset	Combined Dataset (UFPE and Unicamp)
Weights	Kaiming Initialization (HE et al., 2015)
Optimizer	Adam
Learning Rate	1×10^{-3}
Observation	Data Augmentation with a multiplier equal to 3 following the strategy outlined on subsection 4.2.1.

Source: The Author (2024).

In this training, the dataset was expanded by a factor of 3, resulting in 52,128 training images. This increase aimed to enhance the model's exposure to a broader variety of cases, including potential edge cases.

4.8.3 Results

Table 11 summarizes the key performance metrics for the model retrained with data augmentation, evaluated on both the Unicamp and UFPE datasets. The augmented training set, which expanded the data volume by a factor of 3 to a total of 52,128 images, aimed to improve model resilience to challenging cases and diverse imaging conditions.

For the Unicamp dataset, the augmented model achieved an MAE of 3.24 $\pm$ 0.18 years and an MSE of 20.61 $\pm$ 0.26 years^2, with an R^2 score of 0.947 and an explained variance of 0.948. The median absolute error for this dataset was 2.16 years, with an IQR of 3.86 years, indicating a consistent distribution of errors.

On the UFPE dataset, the model achieved an MAE of 3.28 $\pm$ 0.20 years and an MSE of 21.39 $\pm$ 0.28 years^2, with an R^2 of 0.948 and an explained variance of 0.949. The median absolute error for this dataset was 2.10 years, with an IQR of 4.16 years, suggesting that the

Table 11 – Experiment 4 Results: Retraining Model with Data Augmentation

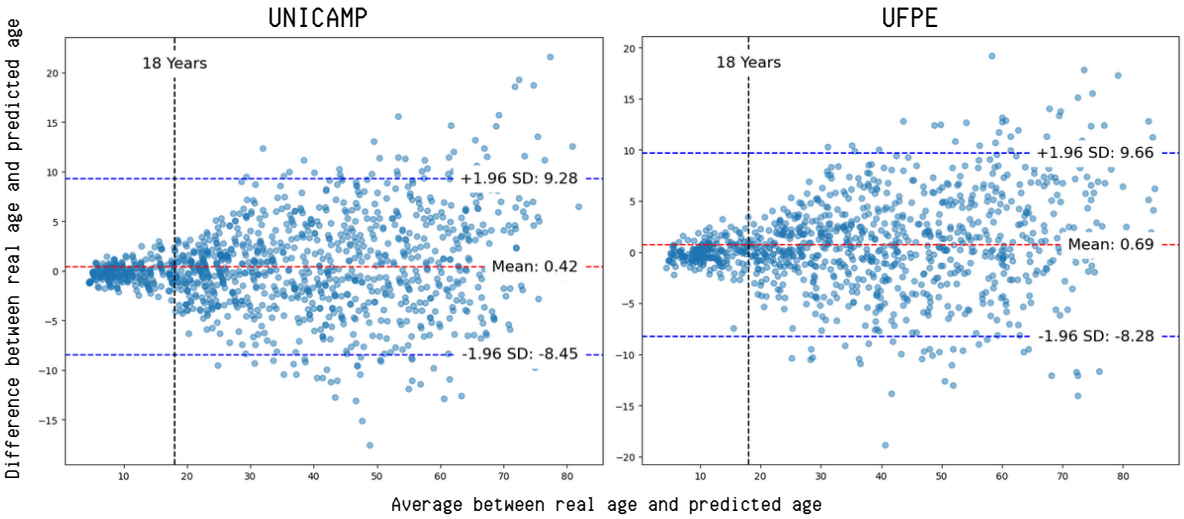
Performance Metrics for Augmented Retrained Model		
	Unicamp Dataset	**UFPE Dataset**
MAE	3.241 $\pm$ 0.182 years	3.284 $\pm$ 0.202 years
MSE	20.614 $\pm$ 0.259 years^2	21.395 $\pm$ 0.283 years^2
Median Absolute Error	2.16 years	2.10 years
IQR of Absolute Error	3.86 years	4.16 years
R^2	0.947	0.948
Explained Variance	0.948	0.949
T-statistic	3.16	4.76
P-value	0.0016	2.17×10^{-6}

Source: The Author (2024). Summary of key performance metrics for Experiment 4, where the model was trained with data augmentation on a combined dataset.

model maintained strong performance across different age ranges and patient conditions.

Although most metrics presented a slight improvement, the P-value indicates that there is statistical evidence that the predictions and the actual ages do not follow the same distribution. This suggests that the model's enhancement to marginally address outliers introduced a greater degree of prediction bias.

Figure 16 – Bland-Altman Plot for Experiment 4



Source: The Author (2024). Comparison of results for the re-trained augmented model across the Unicamp and UFPE datasets.

The Bland-Altman analysis depicted in Figure 16 reveals that, while the augmented model was trained to reduce outliers, this approach did not yield the expected improvement. Although there was a slight reduction in the frequency of extreme errors, the persistence of

high-magnitude outliers suggests that data augmentation alone was insufficient to fully address these cases. Consequently, this trade-off, evident from the P-value analysis, indicates that the augmentation strategy did not significantly enhance the model's ability to generalize across the diverse sample populations in Unicamp and UFPE datasets.

4.8.4 Analysis and Discussion

The results from Experiment 4, where the model was trained with data augmentation, show that while the approach marginally improved some metrics, it did not fully address the presence of high-magnitude outliers. The augmentation increased the data diversity by simulating variations in imaging conditions and patient profiles, resulting in an expanded training set of 52,128 images. This expansion aimed to enhance the model's ability to handle edge cases and challenging inputs.

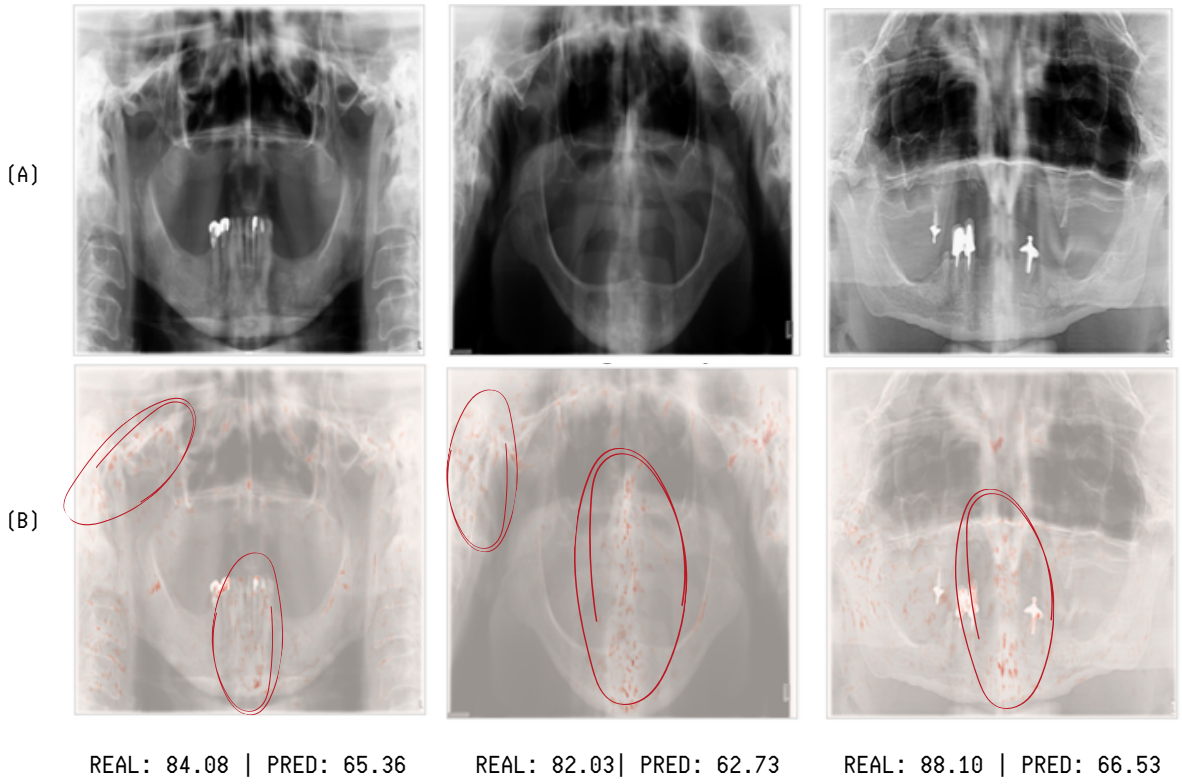
For both datasets, the MAE and MSE values were slightly reduced compared to prior experiments, and the model maintained high R^2 and explained variance values, suggesting a solid overall performance. However, the Bland-Altman plots (Figure 16) indicate that although augmentation helped reduce some extreme errors, it did not eliminate them, nor did it lead to the intended reduction in outlier impact. The T-statistic and P-values further underscore this outcome, showing statistical evidence that the predictions and real ages do not follow the same distribution.

It is important to highlight that the augmentation process, applied uniformly to all images without any targeted filtering, exacerbated the bias related to the imbalance of age groups in the dataset. The dataset itself is predominantly composed of younger and middle-aged adults, while older adults and children are slightly underrepresented. This augmentation inadvertently amplified the over-representation of the majority age groups, leading to improved performance for these groups but increased errors for the minority groups.

As shown in Figure 17, the model consistently underestimates the age of elderly patients—those clearly within an advanced age range, representing a much smaller dataset volume. This tendency to underestimate, particularly evident in high-age patients, points to a significant limitation of the augmentation approach when applied without a stratified strategy to balance the dataset across different age ranges.

On the positive side, the augmentation process did improve the accuracy of predictions for the majority group—namely, young adults and adults—by enhancing the model's robustness

Figure 17 – Experiment 4 - Model Failures for Advanced Age Patients



Source: The Author (2024). Examples of model failures in Experiment 4, focusing on elderly patients where the model significantly underestimates their age. All three cases illustrate the consistent underestimation across different elderly individuals, showing the impact of an unbalanced dataset. Each column presents the original image and integrated gradient visualizations to indicate the regions of focus for the model during prediction.

to variations commonly observed in these populations. However, this improvement came at a cost: the augmentation increased the prediction errors for the minority groups, specifically children and elderly patients, highlighting an inherent trade-off in the approach taken.

These findings indicate that, while data augmentation can be a powerful tool to improve model robustness, it must be carefully tailored to account for the inherent biases within the dataset. In this case, augmentation without specific attention to age group balance led to an unintentional reinforcement of the dataset's original imbalance. Future work should focus on implementing a more targeted augmentation strategy, ensuring balanced representation across all age groups to mitigate these effects and improve generalizability for underrepresented populations.

5 DISCUSSION

The series of experiments conducted in this work provided crucial insights into the performance and generalizability of the age estimation model when exposed to different datasets, training approaches, and augmentation strategies. Each experiment aimed to progressively enhance model performance while addressing regional biases and age-related variabilities. This subsection presents a consolidated discussion of the main findings from Experiments 1 to 4, highlighting the evolution in model capability and its implications for future research.

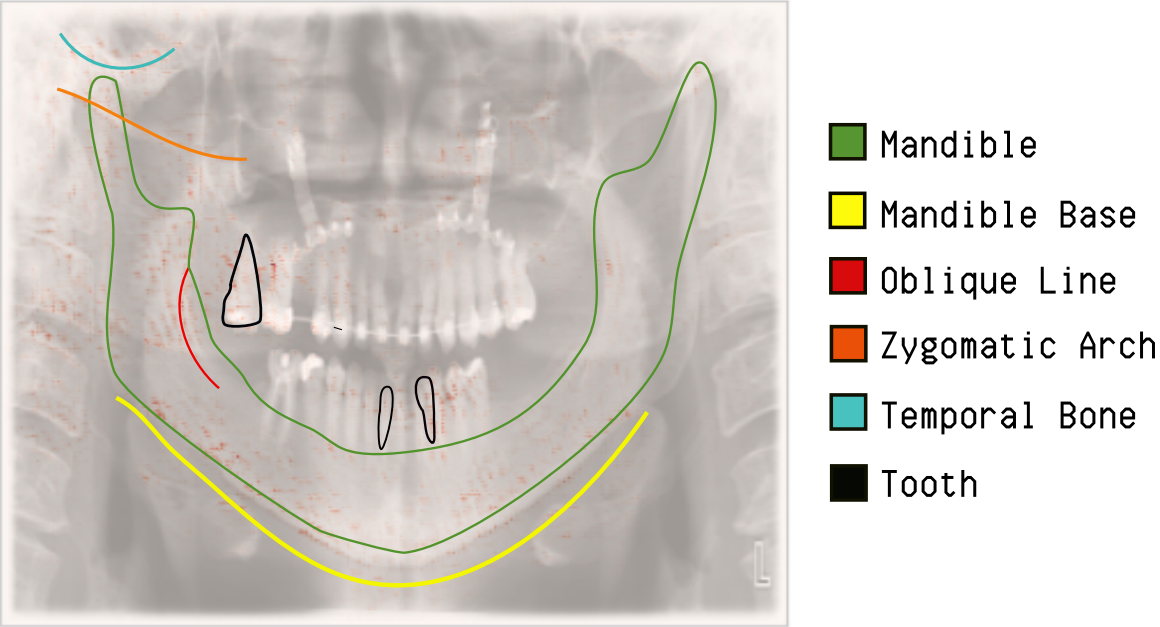
In Experiment 1, the baseline model, trained on the UFPE dataset, was directly applied to the Unicamp dataset without any modifications or fine-tuning. This approach highlighted the initial generalizability limitations, with the model displaying a marked drop in accuracy on the new dataset. The significant decline in metrics such as MAE and R^2 exposed regional biases, indicating that the model's learned features did not adequately represent diverse populations. The Bland-Altman plots confirmed an evident overestimation tendency, particularly for patients above 18 years of age. This experiment served as a critical reference point for understanding the magnitude of regional discrepancies that the subsequent experiments sought to address.

Experiment 2 introduced fine-tuning strategies using two different datasets: Unicamp-only and a combined dataset (UFPE + Unicamp). Fine-tuning using the Unicamp dataset improved the accuracy for this specific region but led to a loss of generalizability, as observed by the poorer performance on the UFPE dataset. Conversely, fine-tuning with the combined dataset yielded a more balanced outcome, maintaining acceptable performance across both datasets. However, even with improved metrics, statistical tests indicated that discrepancies persisted, and the model struggled to generalize perfectly across distinct regions. This experiment emphasized the trade-off between local optimization and broader generalizability.

In Experiment 3, the model was trained from scratch using a combined dataset from both regions. This approach yielded the best overall results in terms of generalizability, with high R^2 scores and minimal bias observed in the Bland-Altman plots. Training from scratch allowed the model to capture diverse age-related features without being constrained by biases present in a pre-trained model. The improvement in statistical alignment between real and predicted ages suggested that a more comprehensive training dataset could effectively address regional biases. However, notable outliers persisted despite significant improvements, highlighting edge cases that remained challenging for the model to predict accurately.

Experiment 4 incorporated data augmentation techniques aimed at further enhancing the model's robustness. Expanding the training dataset with augmented images aimed to introduce more variability and thus improve performance on underrepresented cases. The results showed marginal improvements in MAE and MSE values and helped reduce some of the more extreme errors observed previously. However, the unfiltered augmentation strategy inadvertently exacerbated the dataset's inherent imbalance, leading to worsened predictions for minority groups such as the elderly and young children. Although the approach succeeded in enhancing accuracy for the majority age groups (young adults and adults), it also amplified the errors for the minority populations, revealing a critical limitation of applying augmentation without careful consideration of demographic balance.

Figure 18 – Annotated Regions Evaluated by Model and Specialists



Source: The Author (2024). Visualization highlights the specific dental regions the models and experts evaluated, including the mandible, mandible base, oblique line, zygomatic arch, temporal bone, and teeth. These regions are important markers commonly used in age estimation by dental specialists.

A team of specialists in dental radiology evaluated the model's prediction heatmaps. It confirmed that the regions assessed by the model correspond to key anatomical markers typically used by experts for age estimation, as illustrated in Figure 18. The annotated regions include critical dental structures such as the mandible, mandible base, oblique line, zygomatic arch, temporal bone, and teeth and their pulp chambers. These regions are known to provide essential age-related information, including the degree of dental pulp calcification and the eruption stage of molars, which are commonly used indicators in human evaluations. The model's focus on these markers suggests that it has successfully learned to target biologically relevant

features that align with established practices in dental radiology, enhancing the reliability of its predictions. From these experiments, several key insights emerge:

Generalizability vs. Specificity: The experiments consistently demonstrated a trade-off between optimizing the model for a specific region versus generalizing across diverse datasets. Fine-tuning approaches were beneficial for local adaptation but reduced the model's performance elsewhere. In contrast, retraining from scratch with a diverse dataset enhanced generalizability.

Importance of Regionally Balanced Training Data: The success of the model trained from scratch with a mixed dataset underscores the importance of having a well-balanced training set to capture a broad range of features. This balance helps mitigate regional biases and improve the consistency of predictions across different populations.

Impact of Data Augmentation: The augmentation approach highlighted the risks associated with unfiltered data expansion. While augmentation can introduce helpful variability, it can also unintentionally reinforce biases present in the original dataset, mainly if the augmentation process is not stratified or filtered based on demographic representation.

While each approach explored in these experiments brought distinct advantages, the need for a more refined and balanced training process is evident. Future efforts should focus on developing a targeted augmentation strategy to improve the representation of underrepresented groups, particularly children and elderly patients while ensuring a robust balance between training data diversity and consistency. These findings contribute to building more reliable, generalizable models for age estimation that are applicable across different demographics and regional contexts.

The insights gained from these experiments form a foundation for our conclusions. Each step of this research journey, from the baseline application of pre-trained models to fine-tuning, full retraining, and the incorporation of data augmentation, has provided a deeper understanding of the challenges and opportunities in developing AI models for age estimation that are both accurate and generalizable. Moving forward to the conclusion, we synthesize the overall findings and discuss their broader implications for future research and clinical applications.

6 CONCLUSION

This study evaluated the potential of deep learning models for age estimation using dental radiographs, focusing on datasets from two distinct Brazilian regions, Unicamp and UFPE. To understand the model's generalizability across varied populations within the same country, we employed a series of experiments to explore different training strategies: applying a pre-trained model on new data, fine-tuning with regional and combined datasets, retraining from scratch, and incorporating data augmentation to improve robustness against challenging imaging conditions. These experiments provided valuable insights into how specific training approaches can influence model adaptability and generalizability.

The primary objectives of this study were threefold: first, to assess the model's capacity to generalize across distinct regional populations; second, to improve the generalization of the model by incorporating diverse datasets during training; and third, to investigate the effect of fine-tuning and data augmentation strategies to handle challenging cases effectively.

The first objective was addressed in Experiment 1, which involved applying the baseline model, trained on the UFPE dataset, directly to the Unicamp dataset without any adaptations. This experiment revealed significant limitations in the model's generalization ability, as evidenced by reduced accuracy and noticeable regional biases. These findings highlighted the necessity of regional fine-tuning to enhance model performance in different populations.

The second objective was effectively achieved in Experiment 3, where the model was trained from scratch using a combined dataset from both the UFPE and Unicamp regions. By incorporating diverse datasets during training, the model demonstrated substantial improvements in generalizability, achieving balanced performance across both datasets. This approach was the most effective in minimizing regional biases and ensuring consistent results across distinct Brazilian populations.

The third objective was explored through Experiments 2 and 4, where we implemented fine-tuning and data augmentation strategies. Fine-tuning with regional and combined datasets (Experiment 2) helped improve the model's performance for specific populations, although it presented challenges in maintaining generalizability. Experiment 4 introduced data augmentation techniques to address challenging cases and expand the training set. The results highlighted both the benefits and limitations of these approaches—fine-tuning showed promise in enhancing local accuracy. At the same time, data augmentation improved general robust-

ness but also introduced new biases, particularly affecting minority groups such as children and elderly patients. These experiments collectively contributed to a deeper understanding of the model's adaptability across diverse scenarios.

Quantitative evaluations were supplemented with qualitative assessments using heatmaps, which were reviewed by experts in dental radiology. The experts confirmed that the model focused on clinically relevant features, aligning with established age estimation methods, such as the Cameriere, Demirjian, and Moorrees Methods. This alignment with expert analysis provides confidence in the interpretability of the models trained in this study. This indicates that they learned to rely on similar anatomical markers as human experts, such as pulp chamber dimensions, molar eruption stages, and skeletal structures.

Despite notable advancements, some challenges remain. High-magnitude outliers persisted across all experiments, particularly in cases involving unique dental conditions or suboptimal image quality. These challenging cases highlight the need for more specialized training approaches to improve performance. Additionally, while this study used datasets from two Brazilian regions, a broader demographic representation would enhance the model's ability to generalize across more diverse populations.

Future research could build upon these findings by incorporating broader datasets from Brazil and internationally, establishing a more comprehensive benchmark for model generalizability. This study is a potential foundation for developing such a benchmark, contributing to a dataset aimed at training models for age estimation tasks across varied global populations.

A promising evolution direction might involve advanced data augmentation and dataset balancing techniques, such as adversarial augmentation paired with stratified sampling. These methods could enhance the model's ability to handle underrepresented groups more effectively and reduce prediction bias for minority age categories.

Further investigation could also focus on leveraging more modern network architectures like Vision Transformers or Swin Transformers. With an expanding dataset volume, these architectures, known for their enhanced capability to handle complex feature representations, might prove more effective in addressing variability across patient profiles.

Exploring advanced approaches like anomaly detection, specialized loss functions, or domain adaptation could also significantly enhance model robustness in challenging cases. By addressing the unique characteristics found in certain dental or skeletal features, these methods offer the potential to refine age prediction consistency across diverse populations further.

While this study achieved significant advancements in understanding the challenges and

opportunities of age estimation using panoramic radiographs, certain evaluations could not be performed due to time and resource constraints. These evaluations are nonetheless crucial for further improving the robustness, accuracy, and applicability of AI models in this domain. Future research should also prioritize these directions to address key gaps and refine existing methodologies.

One such direction involves evaluating the impact of using higher-resolution images, which could significantly enhance model accuracy by capturing finer details of dental and skeletal structures. Higher-resolution images would allow the model to better identify subtle anatomical features, such as micro-fractures or minor variations in bone density, which are often critical for precise age estimation. However, this would also require addressing the computational trade-offs involved, such as increased memory usage and training times. Researchers should consider benchmarking models trained with standard and high-resolution images to quantify these impacts.

Another promising approach could involve partitioning the dataset into three or four distinct regions and training an initial classifier to segment images based on their region of origin. This regional segmentation would enable the development of more specialized models tailored to the unique demographic, anatomical, and imaging characteristics of each region, potentially reducing the biases introduced by training on aggregated data.

Studying the latent representations or feature vectors generated by the network could provide deeper insights into the features the model learns to associate with age estimation. This analysis could help identify whether the model relies on clinically relevant attributes or artifacts, guiding further refinement of the network's architecture or training process. Techniques such as t-SNE or UMAP could be employed to visualize the latent space and assess clustering patterns related to age or other demographic factors.

Furthermore, systematically evaluating how the size of the training dataset influences model performance could provide critical insights into data scalability. By incrementally increasing the training set size and observing corresponding improvements in accuracy, it would be possible to estimate the marginal utility of additional data and predict the saturation point beyond which adding more images yields diminishing returns. This analysis would also inform resource allocation for future data collection efforts.

Finally, conducting a direct comparison between the model‘s performance and human experts analyzing the dataset could offer a valuable benchmark. This comparison would highlight the areas where the model matches or exceeds human performance, as well as cases where it

struggles. Such evaluations could also identify scenarios where combining AI predictions with expert oversight could yield better outcomes than either approach alone.

In conclusion, this study demonstrated the feasibility and promise of using deep learning models for age estimation with dental radiographs, highlighting key considerations for achieving generalizability across diverse populations. Our findings underscore the importance of a diverse, well-structured training dataset and specialized training approaches to handle unique challenges. Leveraging these insights in future research can lead to more robust and generalizable models, enhancing the accuracy and applicability of AI-driven age estimation in real-world clinical settings.

REFERENCES

ALKASS, K.; BUCHHOLZ, B. A.; OHTANI, S.; YAMAMOTO, T.; DRUID, H.; SPALDING, K. L. Age estimation in forensic sciences: application of combined aspartic acid racemization and radiocarbon analysis. *Molecular & cellular proteomics*, ASBMB, v. 9, n. 5, p. 1022–1030, 2010.

AVUÇLU, E.; BAŞÇIFTÇI, F. New approaches to determine age and gender in image processing techniques using multilayer perceptron neural network. *Applied Soft Computing*, Elsevier, v. 70, p. 157–168, 2018.

BANAR, N.; BERTELS, J.; LAURENT, F.; BOEDI, R. M.; TOBEL, J. D.; THEVISSEN, P.; VANDERMEULEN, D. Towards fully automated third molar development staging in panoramic radiographs. *International Journal of Legal Medicine*, Springer, v. 134, p. 1831–1841, 2020.

BANG, G.; RAMM, E. Determination of age in humans from root dentin transparency. *Acta Odontologica Scandinavica*, Taylor & Francis, v. 28, n. 1, p. 3–35, 1970.

BEKAERT, B.; KAMALANDUA, A.; ZAPICO, S. C.; VOORDE, W. Van de; DECORTE, R. Improved age determination of blood and teeth samples using a selected set of dna methylation markers. *Epigenetics*, Taylor & Francis, v. 10, n. 10, p. 922–930, 2015.

BENJAVONGKULCHAI, S.; PITTAYAPAT, P. Age estimation methods using hand and wrist radiographs in a group of contemporary thais. *Forensic science international*, Elsevier, v. 287, p. 218–e1, 2018.

BOEDI, R. M.; BANAR, N.; TOBEL, J. D.; BERTELS, J.; VANDERMEULEN, D.; THEVISSEN, P. W. Effect of lower third molar segmentations on automated tooth development staging using a convolutional neural network. *Journal of Forensic Sciences*, Wiley Online Library, v. 65, n. 2, p. 481–486, 2020.

CAMERIERE, R.; BROGI, G.; FERRANTE, L.; MIRTELLA, D.; VULTAGGIO, C.; CINGOLANI, M.; FORNACIARI, G. Reliability in age determination by pulp/tooth ratio in upper canines in skeletal remains. *Journal of forensic sciences*, Wiley Online Library, v. 51, n. 4, p. 861–864, 2006.

CAMERIERE, R.; CINGOLANI, M.; FERRANTE, L. Variations in pulp/tooth area ratio as an indicator of age: a preliminary study. *Journal of forensic sciences*, ASTM International, v. 49, n. 2, p. JFS2003259, 2004.

CAMERIERE, R.; FERRANTE, L.; CINGOLANI, M. Age estimation in children by measurement of open apices in teeth. *International journal of legal medicine*, Springer, v. 120, p. 49–52, 2006.

CAMERIERE, R.; LUCA, S. D.; ALEMÁN, I.; FERRANTE, L.; CINGOLANI, M. Age estimation by pulp/tooth ratio in lower premolars by orthopantomography. *Forensic science international*, Elsevier, v. 214, n. 1-3, p. 105–112, 2012.

CHEN, S.; LV, Y.; WANG, D.; YU, X. Aspartic acid racemization in dentin of the third molar for age estimation of the chaoshan population in south china. *Forensic science international*, Elsevier, v. 266, p. 234–238, 2016.

DALITZ, G. Age determination of adult human remains by teeth examination. *Journal of the Forensic Science Society*, Elsevier, v. 3, n. 1, p. 11–21, 1962.

DEMIRJIAN, A.; GOLDSTEIN, H.; TANNER, J. M. A new system of dental age assessment. *Human biology*, JSTOR, p. 211–227, 1973.

DOSOVITSKIY, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

ELFAWAL, M. A.; ALQATTAN, S. I.; GHALLAB, N. A. Racemization of aspartic acid in root dentin as a tool for age estimation in a kuwaiti population. *Medicine, Science and the Law*, SAGE Publications Sage UK: London, England, v. 55, n. 1, p. 22–29, 2015.

FARGES, J.-C.; ALLIOT-LICHT, B.; RENARD, E.; DUCRET, M.; GAUDIN, A.; SMITH, A. J.; COOPER, P. R. Dental pulp defence and repair mechanisms in dental caries. *Mediators of inflammation*, Wiley Online Library, v. 2015, n. 1, p. 230251, 2015.

FERNANDES, M. M.; TINOCO, R. L. R.; BRAGANCA, D. P. P. de; LIMA, S. H. R. de; JUNIOR, L. F.; JUNIOR, E. D. Age estimation by measurements of developing teeth: accuracy of cameriere's method on a brazilian sample. *Journal of forensic sciences*, Wiley Online Library, v. 56, n. 6, p. 1616–1619, 2011.

GALIBOURG, A.; CUSSAT-BLANC, S.; DUMONCEL, J.; TELMON, N.; MONSARRAT, P.; MARET, D. Comparison of different machine learning approaches to predict dental age using demirjian's staging approach. *International Journal of Legal Medicine*, Springer, v. 135, p. 665–675, 2021.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 1026–1034.

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*. [S.l.: s.n.], 2016. p. 770–778.

HOU, W.; LIU, L.; GAO, J.; ZHU, A.; PAN, K.; SUN, H.; ZHENG, N. Exploring effective dnn models for forensic age estimation based on panoramic radiograph images. In: IEEE. *2021 international joint conference on neural networks (IJCNN)*. [S.l.], 2021. p. 1–8.

HU, Z.; ZHANG, J.; GE, Y. Handling vanishing gradient problem using artificial derivative. *IEEE Access*, v. 9, p. 22371–22377, 2021.

KIM, S.; LEE, Y.-H.; NOH, Y.-K.; PARK, F. C.; AUH, Q.-S. Age-group determination of living individuals using first molar images based on artificial intelligence. *Scientific reports*, Nature Publishing Group UK London, v. 11, n. 1, p. 1073, 2021.

LIU, Z.; TAN, Y.; HE, Q.; XIAO, Y. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, v. 32, n. 7, p. 4486–4497, 2021.

MÁRQUEZ-RUIZ, A. B.; GONZÁLEZ-HERRERA, L.; LUNA, J. d. D.; VALENZUELA, A. Dna methylation levels and telomere length in human teeth: usefulness for age estimation. *International Journal of Legal Medicine*, Springer, v. 134, p. 451–459, 2020.

MOORREES, C. F.; FANNING, E. A.; JR, E. E. H. Age variation of formation stages for ten permanent teeth. *Journal of dental research*, SAGE Publications Sage CA: Los Angeles, CA, v. 42, n. 6, p. 1490–1502, 1963.

MORSE, D. R.; ESPOSITO, J. V.; SCHOOR, R. S.; WILLIAMS, F. L.; FURST, M. L. A review of aging of dental components and a retrospective radiographic study of aging of the dental pulp and dentin in normal teeth. *Quintessence International*, v. 22, n. 9, 1991.

MOURA, R. Rodrigues de; COELHO, A. V. C.; BALBINO, V. de Q.; CROVELLA, S.; BRANDÃO, L. A. C. Meta-analysis of brazilian genetic admixture and comparison with other latin america countries. *American Journal of Human Biology*, Wiley Online Library, v. 27, n. 5, p. 674–680, 2015.

MUMUNI, A.; MUMUNI, F. Data augmentation: A comprehensive survey of modern approaches. *Array*, Elsevier, p. 100258, 2022.

NOLLA, C. M. et al. *The development of permanent teeth*. Tese (Doutorado) — University of Michigan Ann Arbor, 1952.

PURANIK, M.; PRIYADARSHINI, C.; UMA, S. R. Dental age estimation methods: A review. *International Journal of Advanced Health Sciences*, v. 1, p. 19–25, 04 2015.

RAJKUMARI, S.; NIRMAL, M.; SUNIL, P.; SMITH, A. A. Estimation of age using aspartic acid racemisation in human dentin in indian population. *Forensic science international*, Elsevier, v. 228, n. 1-3, p. 38–41, 2013.

RICUCCI, D.; LOGHIN, S.; LIN, L. M.; SPÅNGBERG, L. S.; TAY, F. R. Is hard tissue formation in the dental pulp after the death of the primary odontoblasts a regenerative or a reparative process? *Journal of dentistry*, Elsevier, v. 42, n. 9, p. 1156–1170, 2014.

SANTOSH, K.; PRADEEP, N.; GOEL, V.; RANJAN, R.; PANDEY, E.; SHUKLA, P. K.; NUAGAH, S. J. et al. Machine learning techniques for human age and gender identification based on teeth x-ray images. *Journal of Healthcare Engineering*, Hindawi, v. 2022, 2022.

SHEN, S.; LIU, Z.; WANG, J.; FAN, L.; JI, F.; TAO, J. Machine learning assisted cameriere method for dental age estimation. *BMC Oral Health*, BioMed Central, v. 21, n. 1, p. 1–10, 2021.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

SPALDING, K. L.; BUCHHOLZ, B. A.; BERGMAN, L.-E.; DRUID, H.; FRISÉN, J. Age written in teeth by nuclear tests. *Nature*, Nature Publishing Group UK London, v. 437, n. 7057, p. 333–334, 2005.

ŠTEPANOVSKỲ, M.; IBROVÁ, A.; BUK, Z.; VELEMÍNSKÁ, J. Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods. *Forensic science international*, Elsevier, v. 279, p. 72–82, 2017.

SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V.; ALEMI, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Proceedings of the AAAI conference on artificial intelligence*. [S.l.: s.n.], 2017. v. 31, n. 1.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. *International conference on machine learning*. [S.l.], 2019. p. 6105–6114.

TOBEL, J. D.; RADESH, P.; VANDERMEULEN, D.; THEVISSEN, P. W. An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study. *The Journal of forensic odonto-stomatology*, International Organization of Forensic Odonto-Stomatology, v. 35, n. 2, p. 42, 2017.

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 transactions on medical imaging*, IEEE, v. 39, n. 7, p. 2374–2384, 2020.

ZABOROWICZ, K.; GARBOWSKI, T.; BIEDZIAK, B.; ZABOROWICZ, M. Robust estimation of the chronological age of children and adolescents using tooth geometry indicators and pod-gp. *International Journal of Environmental Research and Public Health*, MDPI, v. 19, n. 5, p. 2952, 2022.