UNIVERSITI TEKNOLOGI MARA

DEVELOPMENT AND VALIDATION OF AN AUTOMATED DEEP LEARNING-BASED CERVICAL VERTEBRAL MATURATION CLASSIFICATION SYSTEM USING OBJECT DETECTION AND MORPHOLOGICAL ANALYSIS

NORAINA HAFIZAN BINTI NORMAN

PhD

June 2026

UNIVERSITI TEKNOLOGI MARA

DEVELOPMENT AND VALIDATION OF AN AUTOMATED DEEP LEARNING-BASED CERVICAL VERTEBRAL MATURATION CLASSIFICATION SYSTEM USING OBJECT DETECTION AND MORPHOLOGICAL ANALYSIS

NORAINA HAFIZAN BINTI NORMAN

Thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Dentistry)

Faculty of Dentistry

June 2026

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 15 April 2026 to conduct the final examination of Noraina Hafizan Binti Norman on her Doctors of Philosophy thesis entitled "Development and Validation of an Automated Deep Learning-Based Cervical Vertebral Maturation Classification System using Object Detection and Morphological Analysis" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiners recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Farha Ariffin, PhD
Associate Professor
Faculty of Dentistry
Universiti Teknologi MARA
(Chairman)

Maryati Md Dasor, MSc
Associate Professor
Faculty of Dentistry
Universiti Teknologi MARA
(Internal Examiner)

Padhraig Fleming, PhD
Professor
School of Dental Science
Dublin Dental University Hospital
Trinity College, Ireland
(External Examiner)

PROFESSOR DR HJH ZURAEDA IBRAHIM
Dean
Institute of Postgraduates Studies
Universiti Teknologi MARA
Date: 30 June 2026

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Noraina Hafizan Binti Norman
Student ID. No.	:	2022640566
Programme	:	Doctor of Philosophy (Dentistry) – DS950
Faculty	:	Dentistry
Thesis Title	:	Development and Validation of an Automated Deep Learning-Based Cervical Vertebral Maturation Classification System using Object Detection and Morphological Analysis
Signature of Student	:	
Date	:	June 2026

ABSTRACT

Accurate assessment of skeletal maturity is fundamental in orthodontics, as treatment timing during growth has a decisive influence on skeletal response, treatment efficiency, and long-term stability. The cervical vertebral maturation (CVM) method provides a practical, radiation-free approach to skeletal maturity assessment using routine lateral cephalograms; however, its clinical utility remains constrained by interpretive subjectivity, transitional morphological ambiguity, and inconsistent reproducibility, particularly at pubertal boundary stages. Although artificial intelligence (AI) has been proposed as a means to standardise CVM staging and support growth-based orthodontic decision-making, progress has been limited by the absence of robust, well-calibrated ground truth datasets and by automated systems that insufficiently reflect clinical reasoning. This study aimed to develop and validate a fully automated CVM classification system that integrates object detection and morphology-driven analysis using deep convolutional neural network architectures, and to benchmark its diagnostic performance against calibrated expert consensus. Recognising that annotation quality is a critical determinant of AI performance, a structured multi-expert calibration and consensus framework was implemented to establish reproducible reference labels. Under these calibrated criteria, 1,047 anonymised lateral cephalograms from patients aged 6 to 17 years were annotated into six CVM stages by two orthodontic experts. The dataset was partitioned into training (80%), validation (10%), and test (10%) subsets. A two-stage automated pipeline was developed, combining YOLOv8 for precise localisation of the second to fourth cervical vertebrae (C2–C4) with LightGBM for classification based on extracted morphological features. This design decoupled anatomical localisation from stage classification, enabling targeted analysis of biologically relevant structures while maintaining model interpretability. Model performance was evaluated using accuracy, precision, recall, F1-score, and agreement-based metrics. Inclusive accuracy and unweighted kappa (κ) were employed to reflect clinically acceptable adjacent-stage variability, while intraclass correlation coefficients were used to quantify intra-observer reliability. The proposed system demonstrated high diagnostic performance, achieving expert-comparable strict classification accuracy and improved inclusive accuracy consistent with clinical staging tolerance. Agreement with expert consensus was substantial to near-perfect across performance metrics, particularly for morphologically distinct pubertal and post-pubertal stages. Residual misclassifications were predominantly confined to biologically transitional CVM stages, reflecting intrinsic morphological continuity rather than systematic algorithmic error. External validation further confirmed patterned behavioural consistency with expert staging and yielded interpretable morphological cues that aligned with established clinical reasoning. Taken together, these findings indicate that reliable, automated CVM staging is achievable when AI development is grounded in expert-calibrated annotation and biologically informed model design. The proposed framework demonstrates the potential for CVM assessment to become more reproducible and clinically consistent through integration of expert-calibrated annotation and AI-assisted analysis. By integrating consensus-based ground truth with object detection and morphology-driven classification, this study demonstrates the translational potential of AI to enhance consistency and reliability in skeletal maturity assessment within orthodontic practice.

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to Allah SWT for granting me the strength, patience, and perseverance to complete this research journey. Without His blessings and guidance, none of this would have been possible.

I owe my sincerest appreciation to my supervisor, Professor Dr. Mohd Yusmiaidil Putera Mohd Yusof, for being the mastermind who connected the dots through his network and for his invaluable guidance, constructive feedback, and continuous encouragement throughout this study. His wisdom and patience have been a true compass in navigating this academic voyage. I am equally grateful to my co-supervisors, Associate Professor Dr. Marshima Mohd Rosli and Associate Professor Dr. Nagham Abdullah Al-Jaf, whose expertise, advice, and critical insights in orthodontics and computer science sharpened both the scope and the soul of this research. Special thanks go to my coresearcher, Dr. Norhasmira Mohammad, who was more than a collaborator, a teammate in late-night data debates, a sounding board when methods needed rethinking, and a companion in keeping perspective when the AI proved unpredictable. My heartfelt gratitude extends to Dr. Mohd Hanif Abdul Gaus and the outstanding technical team at our medical AI industry partner for their brilliance in model development and deployment, and for proving that teamwork truly makes the (neural) dream work.

To my research assistants, Karmila Syuhada and Nur Ain Sofeea, for always being ready to travel at a moment's notice and for tackling last-minute assignments with dedication and grace. My heartfelt thanks also go to my colleagues at the Centre of Studies for Paediatric Dentistry and Orthodontics for holding the fort in my absence. To my best friend, Melati Mahmud, and my gal pals, thank you for always keeping tabs on me, making sure I was still alive (and mostly sane) during this journey. This research was never a solo marathon but a relay race, and I feel truly blessed to have had all of you running alongside me as part of the team.

On the home front, my deepest appreciation goes to my children, Hana Laiqa, Lana Zakia, Dana Sadia, and Rumi Ismet. You have been the greatest source of joy and strength, showing patience and understanding whenever I needed to focus on my work. You reminded me daily why perseverance matters, and why moments of family love are more valuable than any academic milestone.

And to my husband and best friend, Mohamed Azim Intizam, my rock, my motivator, and my steady companion through countless last-minute work, thank you for holding everything together when I was juggling a million things. You stood by me unfailingly, cared for our children with such love, and made sure I had the space and time to burn the midnight oil. This milestone is as much yours as it is mine, and I am forever grateful to you.

Lastly, to my parents, in laws and family members, whose prayers and unwavering belief in me have carried me further than words can ever capture, this accomplishment is a true reflection of your love and sacrifices. Alhamdulillah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations

2D	Two-dimension
AI	Artificial Intelligence
ALARP	As low as reasonably practicable
ANNs	Artificial Neural Networks
AUC	Area Under the Curve
CBCT	Cone Beam Computed Tomography
CI	Confidence Interval
CLAIM	Checklist for Artificial Intelligence in Medical Imaging
CNNs	Convolutional Neural Networks
CQ	Category Quartile
CSV	Comma-Separated Values
CVM	Cervical Vertebral Maturation
DCNN	Deep Convolutional Neural Network
DM	Dental Maturation
DPTs	Dental Panoramic Tomographs
FN	False negative
FP	False positive
GCF	Gingival Crevicular Fluid
GPU	Graphics Processing Unit
Grad CAM	Gradient-weighted Class Activation Mapping
HWM	Hand-Wrist Method
ICC	Intraclass Correlation Coefficient
IF	Impact Factor

JCI	Journal Citation Indicator
LightGBM	Light Gradient Boosting Machine
MDPI	Multidisciplinary Digital Publishing Institute
MSU	Management and Science University
NA	Not available
PGS	Pubertal Growth Spurt
PHV	Peak Height Velocity
QCVM	Quasi-continuous CVM Representations
RAM	Random Access Memory
ReLU	Rectified Linear Unit
RF	Random Forest
RME	Rapid Maxillary Expansion
ROC	Receiver Operating Characteristic
ROI	Region of Interest
ROS	Random Oversampling
SJR	SCImago Journal Rank
SNIP	Source Normalised Impact per Paper
SMI	Skeletal Maturation Indicators
SMOTE	Synthetic Minority Over-sampling Technique
SVMs	Support vector machines
TN	True negative
TP	True positive
UIAM	Universiti Islam Antarabangsa Malaysia
UiTM	Universiti Teknologi MARA
UKM	Universiti Kebangsaan Malaysia

UM	Universiti Malaya
USIM	Universiti Sains Islam Malaysia
USM	Universiti Sains Malaysia
XGBoost	Extreme Gradient Boosting

CHAPTER 1
INTRODUCTION

1.1 Background

The accurate assessment of skeletal maturity is a cornerstone of orthodontic diagnosis and treatment planning. Successful orthodontic and dentofacial orthopaedic interventions often depend on optimally timing treatment to coincide with periods of accelerated growth. For example, functional appliances intended to modify mandibular growth may demonstrate greater skeletal effects when treatment is initiated around the pubertal growth spurt, although individual response variability and the limitations of growth indicators must also be recognised (Mandall et al., 2022). When treatment is delivered outside periods of active growth, the skeletal contribution to treatment may be reduced, potentially increasing reliance on dentoalveolar compensation or combined orthodontic-surgical approaches in selected cases (Mandall et al., 2016a).

The ability to determine an individual's skeletal maturity and remaining growth potential therefore plays a central role in identifying the ideal timing for correcting skeletal discrepancies, guiding treatment planning, and enhancing the accuracy of age estimation. Similarly, the decision to delay orthognathic surgery until growth completion requires reliable assessment of skeletal maturity (Tulloch et al., 2004).

Chronological age and skeletal age are two different concepts in the literature. Numerous studies have demonstrated that chronological age is not a valid indicator of skeletal maturation (Cavallo et al., 2021; Hagg & Pancherz, 1988). Since chronological age and skeletal age frequently do not coincide, individual skeletal development can vary widely (Miltenburg Caspersen & Sonnesen, 2020). Nevertheless, skeletal age, when assessed using cervical vertebral maturation (CVM) stage classification from lateral cephalometric radiographs, may still offer relative predictability of maturation patterns and therefore remains a parameter of interest in research examining growth timing and developmental progression in growing patients (Abdulla Alkhal et al., 2008; Baccetti et al., 2006)

There are several ways of predicting the optimum growth stage which involves assessing the skeletal maturity of a growing child. A well-known method for determining the dentoskeletal maturation stage, growth velocity, and prediction of

residual growth remaining is to employ radiographic assessments (Cericato et al., 2015; Magat & Ozcan, 2022). Over the past century, several methods have been developed to evaluate skeletal maturation. Hand–wrist radiographs, based on ossification events in the small bones of the hand and wrist, were historically considered the gold standard (Fishman, 1982; Pyle & Greulich, 1959). These methods provide detailed information on growth potential but require additional radiographic exposure, which introduces unnecessary radiation in cases where lateral cephalograms are already available for orthodontic diagnosis (Cavallo et al., 2021). Other approaches have included chronological age, tooth development and eruption times, and secondary sexual characteristics, but these indicators are limited by wide inter-individual variability and relatively poor correlation with true biological growth (Abdulla Alkhal et al., 2008; Hagg & Pancherz, 1988; Miltenburg Caspersen & Sonnesen, 2020).

In response to the limitations of alternative indicators, CVM analysis was introduced as a practical method for estimating skeletal maturity using standard lateral cephalograms (Perinetti & Contardo, 2017). The availability of standard lateral cephalograms as a routine component of an orthodontic study record has contributed to their usage (Szemraj et al., 2018). The CVM method evaluates morphological changes in the cervical vertebrae, particularly C2, C3, and C4, across six developmental stages (Perinetti & Contardo, 2017). As the vertebrae transition from trapezoidal to rectangular to square shapes, and as inferior border concavities become more pronounced, these features are correlated with the onset, peak, and cessation of pubertal growth (Franchi et al., 2021; Perinetti et al., 2014) The appeal of CVM lies in its integration into routine orthodontic workflows without requiring additional imaging, thereby avoiding unnecessary radiation exposure since the cephalograms are part of the diagnostic process (McNamara & Franchi, 2018).

Despite its clinical utility, CVM staging is not without limitations. The morphological changes between consecutive stages of vertebral maturation are often subtle, particularly during transitional phases such as CVM Stage III and CVM Stage IV(Morris et al., 2019; Sohrabi et al., 2016; Zhang et al., 2020). These stages are critical because they correspond to the acceleration and deceleration phases of growth, yet they present the greatest diagnostic difficulty. Numerous studies have reported substantial intra- and inter-observer variability in CVM classification, even among experienced orthodontists (Nestman et al., 2011; Perinetti et al., 2012). Discrepancies in identifying concavity development or categorising vertebral shape have raised concerns about

reproducibility and reliability (Franchi et al., 2021). This may point to the challenge of conducting an accurate CVM assessment, which can be overcome by a qualitative assessment method or intelligent systems. Different technologies of radiological image interpretation have been used in dentistry over time (Schlégl et al., 2022).

In recent years, artificial intelligence (AI) describes a machine's capacity to carry out cognitive functions akin to those of human intelligence. Deep learning is a subset of AI that processes complex data using a hierarchical network of algorithms (Makaremi et al., 2019). Deep learning models, particularly convolutional neural networks (CNNs), have shown promise in medical image analysis tasks such as lesion detection, radiographic segmentation, and disease classification (Esteva et al., 2017). Dentistry and orthodontics have also benefitted from AI innovations, including skeletal classification, landmark identification on lateral cephalometry, and in determining the thresholds for orthognathic intervention (Kök et al., 2019; J. M. Lee et al., 2024). Within this context, researchers have begun applying AI to automate CVM staging. The majority of these studies used semi-automated algorithm to process manually created features, including landmarks. In addition, early attempts demonstrated moderate accuracy but were hampered by small datasets, lack of standardised ground truth, and poor generalisability (Amasya, Cesur, et al., 2020; Seo et al., 2021). Newer advances in the AI field are exploring fully automated systems to eliminate the need for human landmark identification which may be prone to internal and external errors (Atici et al., 2022).

1.2 Problem Statement

CVM analysis offers several advantages that support its widespread use in orthodontics. A key advantage is that CVM staging can be performed using lateral cephalometric radiographs, which are already routinely obtained for orthodontic diagnosis and treatment planning. This avoids the need for additional imaging, prevents unnecessary radiation exposure, and provides a practical and cost effective method for evaluating skeletal maturity (Baccetti et al., 2006; Hassel & Farman, 1995). Numerous studies have demonstrated significant correlation between CVM stages and hand–wrist maturation indicators, further supporting its clinical utility as a non-invasive alternative for assessing growth (Cunha et al., 2018; D. W. Kim et al., 2021; Tekın & Cesur Aydın, 2020; Uysal et al., 2006).

Despite these advantages, several limitations diminish the reliability and clinical applicability of CVM assessment. A major challenge is the variability and subjectivity inherent in manual staging (Ali et al., 2024). Studies consistently report only moderate inter observer agreement, particularly at transitional stages where morphological changes are subtle and difficult to distinguish. Reported Cohen's kappa values typically range from 0.60 to 0.75, indicating substantial yet imperfect reproducibility (Jaqueira et al., 2010; Rainey et al., 2016). Examiner subjectivity is especially pronounced in evaluating features such as inferior border concavity or distinguishing between rectangular and square vertebral shapes. Poor reproducibility is well documented among non-expert assessors (Echevarría-Sánchez et al., 2020a; Gudhimella et al., 2022). Although training and calibration can improve staging consistency, manual assessment remains highly dependent on experience, and perfect reliability is unlikely to be achieved (Perinetti et al., 2018, 2020).

Limitations also extend to digital and AI based approaches. Standard digital cephalometric analysis software does not incorporate CVM morphological criteria, requiring clinicians to rely on subjective visual assessment. Attempts to automate CVM staging have often used semi-automated methods requiring clinicians to manually identify up to 30 anatomical landmarks. This reliance on manual landmarking introduces further subjectivity and restricts standardisation (Khanagar, Al-Ehaideb, et al., 2021). Fully automated AI models remain limited, partly due to small datasets and the absence of large, well validated image collections (Atici et al., 2022; Mathew et al., 2022). While convolutional neural networks have shown promise in medical imaging, CVM classification remains challenging due to subtle morphological variations, imbalanced class distribution, and limited data (Amasya, Cesur, et al., 2020). Reported validation accuracies typically range between 70 percent and 85 percent, with persistent difficulty distinguishing transitional stages such as CVM Stages III and IV (Amasya, Yildirim, et al., 2020; Atici et al., 2023). Overfitting, poor generalisability, and reliance on single institution datasets further restrict the clinical utility of these models (Atici et al., 2023).

Another significant barrier to progress is the absence of validated ground truth datasets. Many studies rely on a limited number of annotators without calibration, consensus protocols, or structured decision rules (Khanagar, Al-Ehaideb, et al., 2021). Without reliable ground truth labels, AI systems risk learning inconsistent or biased staging criteria, an especially problematic issue given the inherent subjectivity of CVM

interpretation (Flores-Mir et al., 2006; Hassel & Farman, 1995). The lack of reproducible annotation standards continues to be one of the most critical limitations in both manual and AI assisted CVM assessment (Khanagar, Al-ehaideb, et al., 2021).

Collectively, these challenges highlight a critical gap in orthodontic diagnostics. Although CVM remains central to assessing skeletal maturation and timing orthodontic and orthopaedic interventions, neither manual staging nor current AI based systems adequately resolve issues of transitional ambiguity, inter examiner variability, dataset inconsistency, and reproducibility (W. Kazimierczak et al., 2024). A clinically validated, fully automated, AI driven CVM classification model capable of delivering objective, reliable, and scalable staging is urgently needed (Dipalma et al., 2023).

Given that detection of the pubertal growth spurt is one of the most clinically valuable applications of CVM staging, the development of an accurate deep learning model using routine lateral cephalometric radiographs represents an important advancement for evidence based orthodontic care (Mathew et al., 2022). Accordingly, the aim of this study is to develop and evaluate a new deep learning model for automated CVM stage classification using lateral cephalometric radiographs.

1.3 Research Justification

The justification for developing an automated, reliable CVM classification system rests on several interrelated dimensions. From a clinical standpoint, orthodontists require diagnostic tools that reduce variability and enhance confidence in treatment planning. An AI system capable of replicating expert consensus could serve as a valuable decision-support tool (Du et al., 2024). By providing objective and reproducible classifications, such a system could reduce examiner fatigue, minimise errors, and ensure more consistent timing of interventions (Prasad et al., 2022). In regions with limited access to orthodontic specialists, automated CVM staging could expand access to high-quality care and reduce inequities in treatment outcomes (Brickley et al., 1998).

From a research standpoint, establishing a ground truth dataset through multi-expert consensus represents a methodological advancement (Prasad et al., 2022). Reliability science emphasises that reproducibility is foundational for trustworthy inference (Gowri Priya & Thirugnanam, 2024). Without reliable labelling, AI systems cannot be meaningfully validated. By demonstrating that CVM staging can achieve

substantial to near-perfect reproducibility when guided by calibration and consensus protocols, this study contributes to both orthodontic methodology and broader digital health standards (Du et al., 2024).

From a technological standpoint, the integration of object detection with feature-based classification is innovative (Ueda et al., 2023). Most previous AI studies applied CNNs directly to entire radiographs, often introducing noise from irrelevant structures (Kapila et al., 2023). By first localising the cervical vertebrae and then classifying their morphology the proposed system mirrors the diagnostic workflow of clinicians while enhancing both anatomical precision and interpretability (Yousefirizi et al., 2022)

Finally, from a digital health perspective, this study aligns with principles of transparency, reproducibility, and fairness (Tejani et al., 2024a). Recent frameworks emphasise that AI systems in healthcare must not only achieve high accuracy but also provide interpretable outputs and withstand external validation (Yousefirizi et al., 2022). By embedding these principles into the developed AI system, the study situates orthodontics within the broader paradigm of evidence-based AI in medicine (Tejani et al., 2024a).

1.4 Research Aims and Objectives

This research addresses the limitations of manual CVM staging and existing automated approaches by integrating expert calibration with deep learning–based classification.

Aim

The primary aim of this study was to develop and validate a fully automated CVM classification system with reliability comparable to multi-expert consensus.

Objectives

1. To evaluate inter- and intra-observer reliability in CVM classification among multiple orthodontic experts.
2. To establish a calibrated, consensus-based ground truth dataset for CVM staging.
3. To develop a fully automated CVM classification system using a two-stage deep learning architecture.
4. To compare the performance of the automated CVM classification system with conventional visual CVM assessment by calibrated experts.

1.5 Research Questions

The study was guided by the following research questions:

RQ1. How consistent and reliable is CVM staging among orthodontic experts prior to model development?

RQ2. Can structured calibration and multi-expert consensus improve the consistency and reliability of CVM stage labelling to establish a robust ground truth dataset?

RQ3. Can a fully automated deep learning–based system accurately classify CVM stages from digital lateral cephalograms?

RQ4. How does the performance of the automated CVM classification system compare with conventional visual CVM assessment by calibrated orthodontic experts?

1.6 Significance of the Study

This study contributes to orthodontics and medical AI, and clinical decision support at multiple levels.

Clinically, it provides an automated CVM staging system capable of delivering diagnostic accuracy comparable to human experts, with the potential to streamline diagnostic workflows, reduce inter observer variability, and expand access to reliable growth assessment in resource limited environments (Nestman et al., 2011; Yousefi et al., 2023). Accurate identification of growth stages is essential for determining optimal timing of functional and orthopaedic interventions. In Class II correction, treatment initiated during the pubertal growth spurt yields significantly greater skeletal changes, shorter treatment duration, and improved long term stability compared to early or late intervention (Franchi et al., 2000; Hagg & Pancherz, 1988; O'Brien et al., 2003). Several studies have reported greater skeletal contribution and improved treatment efficiency when functional appliance therapy is initiated during periods of active growth, although findings remain influenced by treatment modality, case selection, and growth variability (Baccetti et al., 2005; O'Brien et al., 2009). Delayed or suboptimal treatment timing may reduce the skeletal contribution of growth modification and

increase reliance on dentoalveolar compensation, particularly in moderate to severe skeletal discrepancies (Baccetti et al., 2005; Tulloch et al., 2004). Earlier initiation of maxillary protraction and rapid maxillary expansion has been associated with more favourable skeletal responses in selected growing patients, although long-term stability and individual treatment response remain variable (Franchi et al., 2000; Ngan & Moon, 2015). Although growth timing remains an important consideration in orthodontic treatment planning, the clinical relevance of CVM staging should not be interpreted in isolation. Treatment outcomes are influenced by multiple interacting factors, including treatment modality, severity of skeletal discrepancy, patient compliance, craniofacial growth variability, and clinician experience (Mandall et al., 2016b). Furthermore, limitations inherent to the CVM method itself, including transitional ambiguity and moderate reproducibility reported in previous literature, should be acknowledged when interpreting skeletal maturation status. Accordingly, the present study positions automated CVM assessment as a supportive decision-making tool rather than a definitive predictor of treatment outcome.

Methodologically, this study demonstrates that high reproducibility in CVM staging can be achieved through structured calibration and consensus scoring, strengthening the validity of ground truth labels used for AI development. This addresses a persistent challenge in CVM research, where examiner subjectivity, especially in transitional stages, has limited staging consistency (Flores-Mir et al., 2006; Hassel & Farman, 1995; Perinetti et al., 2012).

Technologically, the research advances AI based skeletal maturity assessment through a hybrid architecture combining object detection with optimised morphological feature classification, balancing anatomical localisation with discriminative modelling power. This contributes to innovation in medical image analysis within dentistry and supports the development of robust, interpretable AI systems aligned with clinical expectations (Shahbazi et al., 2025).

Scientifically and within digital health frameworks, the study aligns with principles of reproducibility, reliability, and interpretability, essential to trustworthy AI in healthcare (World Health Organization, 2021). By situating orthodontic AI within broader validation standards, the study provides a model for developing safe, evidence based clinical decision support systems capable of assisting practitioners in diverse clinical settings, including underserved regions where orthodontic specialists are limited.

Overall, by automating a diagnostically influential step that directly affects treatment timing, skeletal outcomes, and resource utilisation, this study offers meaningful contributions to orthodontic care, medical imaging AI, and equitable access to growth assessment expertise (Shahbazi et al., 2025).

1.7 Scope and Limitations of the Study

This study focuses on developing and validating a fully automated CVM classification system using lateral cephalometric radiographs of patients aged 6–17 years. It encompasses the creation of an anonymised dataset, expert-calibrated CVM staging to establish reliable ground truth, and the construction of an automated framework designed to identify and classify CVM stages. Model performance was evaluated using established diagnostic metrics, including accuracy, kappa statistics, ICC, and ROC analysis benchmarked against expert consensus. The scope is limited to two-dimensional lateral cephalograms and does not include alternative skeletal maturity indicators such as hand–wrist radiographs, CBCT imaging, or biochemical markers.

Despite rigorous methodology, several limitations must be acknowledged. Using a single 2D imaging modality introduces constraints related to superimposition, radiographic variability, and differences in patient positioning. The dataset originates from a limited number of institutions, which may limit generalisability across populations with differing growth patterns. Under-representation of early CVM stages reflects ethical constraints in imaging younger children, resulting in unavoidable class imbalance. Transitional stages (CVM Stages III and IV), characterised by overlapping morphological features, pose inherent challenges for both human and automated classification. External, multi-centre validation, integration of advanced explainable-AI techniques, and real-world clinical deployment remain outside the scope of this study but represent essential directions for future work to enhance robustness and clinical applicability.

CHAPTER 2
LITERATURE REVIEW

2.1 Pubertal Growth Spurt and Clinical Relevance

The pubertal growth spurt (PGS) is a pivotal biological event that marks the most rapid phase of skeletal and somatic development during childhood and adolescence (Abdulla Alkhal et al., 2008; Soegiharto, Moles, et al., 2008). This period is characterised by accelerated longitudinal bone growth, maturation of craniofacial structures, and significant changes in musculoskeletal proportions, all of which have direct implications for orthodontic diagnosis and treatment planning (Hagg & Pancherz, 1988; Mellion et al., 2013). In orthodontics, accurately identifying and utilising the PGS is fundamental because many growth modifying interventions, including functional appliance treatment for Class II malocclusion, orthopaedic maxillary protraction in Class III patients, and transverse expansion procedures, may demonstrate enhanced skeletal responsiveness when treatment is initiated during periods of active pubertal growth (Baccetti et al., 2005; McNamara & Franchi, 2018; Ngan & Moon, 2015). Treatment initiated outside periods of accelerated growth may demonstrate comparatively reduced skeletal effects in some patients (Franchi et al., 1999; Hagg & Pancherz, 1988; Perinetti & Contardo, 2017).

From a clinical perspective, the PGS offers a crucial opportunity to guide or modify facial growth in a biologically efficient and predictable manner. Accurate estimation of the onset, peak, and decline of this growth phase allows clinicians to anticipate mandibular and maxillary growth patterns, optimise the timing of interceptive therapy, refine decisions regarding extraction versus non extraction treatment strategies, and appropriately plan orthopaedic or surgical approaches (Cavallo et al., 2021; Khade et al., 2023; Tulloch et al., 2004). Because chronological age is a poor predictor of individual skeletal maturity, orthodontists have historically relied on alternative indicators, such as hand wrist radiographs, dental maturation stages, physiological maturation markers, and CVM, to obtain a more accurate estimation of pubertal timing and growth potential (Baccetti et al., 2006; Demirjian A, 1973; Hassel & Farman, 1995; Mellion et al., 2013).

Understanding the biological basis, variability, and clinical significance of the PGS therefore remains essential for evidence based orthodontic practice. Utilising periods of active growth may enhance the skeletal component of orthodontic and orthopaedic treatment in selected growing patients.

2.2 Indicators of Skeletal Maturation

A fundamental step in assessing the pubertal growth spurt is selecting a reliable indicator of skeletal and biological maturation. Over the years, clinicians and researchers have recognized that no single measure sufficiently captures the complexity of adolescent growth, largely due to considerable variability in developmental timing among individuals (Cunha et al., 2018; Mellion et al., 2013). As a result, multiple diagnostic approaches have been developed to estimate the timing, intensity, and progression of the pubertal growth spurt, each reflecting different dimensions of physiological development (Baccetti et al., 2005; McNamara & Franchi, 2018). These methods provide orthodontists with essential information for optimizing treatment timing, especially for growth-modifying and interceptive interventions, where even small differences in maturation stage can significantly influence outcomes (Abdulla Alkhal et al., 2008; Hassel & Farman, 1995).

Given the multifactorial nature of adolescent maturation, these diagnostic approaches are commonly categorized into clinical, radiographic, and biochemical indicators, each offering different perspectives on the pubertal growth process. Clinical indicators such as chronological age, dental age, and physiological age provide preliminary information based on observable developmental milestones, although their accuracy is limited by substantial inter-individual variation (Cavallo et al., 2021; Mellion et al., 2013). Radiographic indicators, including dental maturation assessments, hand–wrist evaluation, and CVM analysis, enable more direct appraisal of skeletal development and have demonstrated stronger correlations with peak growth velocity, making them central to contemporary orthodontic diagnostics (Baccetti et al., 2005; Demirjian et al., 1985; Fishman, 1982; Franchi et al., 2000; Hassel & Farman, 1995; McNamara & Franchi, 2018). More recently, biochemical or laboratory-based markers, such as serum and salivary hormonal biomarkers, have been investigated as adjunctive tools for assessing maturation, reflecting the underlying endocrine changes associated with pubertal progression (Gupta et al., 2012; Trehan & Patil, 2021). Collectively, these

categories provide a comprehensive framework for evaluating the pubertal growth spurt and enhancing the precision of growth-related treatment planning in orthodontics.

2.2.1 Clinical Indicators

Clinical indicators represent the earliest and most accessible methods used to estimate the onset and progression of the PGS. Historically, clinicians have relied on observable developmental parameters such as chronological age, physiological maturation, dental development, and anthropometric measurements to approximate an individual's biological growth status (Cavallo et al., 2021). Although these indicators provide useful preliminary information, their accuracy in identifying the precise timing of the PGS is limited due to wide inter-individual variation influenced by genetic, hormonal, nutritional, and environmental factors (Cunha et al., 2018; Mellion et al., 2013).

Chronological age refers to the time elapsed since birth and is the simplest clinical estimate of development. However, its correlation with skeletal maturation is weak, making it an unreliable predictor of peak growth velocity and biological maturity (Abdulla Alkhal et al., 2008). Physiological age provides additional context by considering observable secondary sexual characteristics such as menarche in girls, voice changes in boys, and general pubertal development which reflect underlying endocrine activity. While more biologically relevant than chronological age, these physiological signs remain subjective and may not coincide precisely with the timing of maximum skeletal growth (Cavallo et al., 2021).

Dental age, assessed through tooth eruption patterns or the radiographic evaluation of dental maturation stages, offers another clinical indicator of development. Although commonly used in paediatric dentistry, dental development demonstrates only moderate correlation with skeletal maturation and is therefore of limited value in accurately identifying the PGS (Demirjian et al., 1985; Mellion et al., 2013).

Anthropometric measures such as height velocity, sitting height, leg length, or arm span have also been used as indirect indicators of pubertal growth. Height velocity, in particular, is closely linked to the timing of peak growth; however, it requires serial measurements and is influenced by systemic factors, reducing its practicality and precision in orthodontic settings (Abdulla Alkhal et al., 2008).

Overall, while clinical indicators contribute valuable background information on a patient's developmental status, they do not provide the diagnostic precision required for growth-modifying orthodontic interventions. For this reason, clinicians frequently supplement or replace clinical indicators with more reliable radiographic and biochemical assessments, which better reflect skeletal maturation and the biological processes underlying the pubertal growth spurt (Baccetti et al., 2005; Fishman, 1982; Gupta et al., 2012; Hassel & Farman, 1995; McNamara & Franchi, 2018; Trehan & Patil, 2021).

2.2.2 Biochemical Indicators

Biochemical indicators have been explored as an alternative approach for assessing skeletal maturity, as they reflect physiological changes associated with bone metabolism and pubertal development. Hormones, peptides and enzymes involved in growth undergo characteristic fluctuations during different stages of puberty, and many of these biomolecules diffuse into bodily fluids such as blood, saliva, gingival crevicular fluid (GCF) and urine (De Sanctis et al., 2014; Tripathi et al., 2017). These fluids act as biological mirrors that provide measurable representations of systemic changes related to maturation (Tripathi et al., 2018).

The diagnostic potential of biochemical markers is influenced by the method of sample collection. Serum-based assessment is accurate but requires invasive venepuncture, which may not be ideal for children (Robins, 1994). GCF sampling is a minimally invasive option but demands specialised equipment and is technique sensitive (Perinetti et al., 2011). Saliva offers a highly accessible, non-invasive and child-friendly medium. It contains a wide range of biomolecules, including proteins, peptides, hormones, nucleic acids and electrolytes, originating from both local and systemic sources. This makes saliva a valuable substrate for growth assessment (Malamud, 2011; Perinetti et al., 2011).

Several studies have investigated potential associations between salivary biomarkers and skeletal maturity indicators, although the available evidence remains limited. Most investigations rely on cross-sectional data, and longitudinal evidence on diagnostic validity, reproducibility and prediction accuracy is still insufficient (Khade et al., 2023). As a result, salivary biomarkers are currently regarded as adjunctive rather

than primary tools for predicting growth or determining skeletal maturation in orthodontic treatment planning.

A number of salivary biomarkers have been studied, including alkaline phosphatase, insulin-like growth factor 1, insulin-like growth factor binding protein 3, cortisol, Indian hedgehog and dehydroepiandrosterone sulfate. Among these, salivary insulin-like growth factor 1 appears to be the most promising indicator for skeletal maturity assessment (Khade et al., 2023; Masoud et al., 2008; Perinetti et al., 2012). Its levels show meaningful associations with pubertal stages and may enhance growth prediction when used in conjunction with established clinical or radiographic methods.

Further research is needed to establish the diagnostic value of biochemical markers. Larger longitudinal studies with standardised collection protocols and analytical methods are essential to determine the accuracy, reproducibility and clinical utility of salivary biomarkers in CVM stage assessment (Khade et al., 2023; Tripathi et al., 2018).

2.2.3 Radiographic Indicators

2.2.3.1 Dental Maturation

Dental panoramic tomographs (DPTs) are among the most frequently used radiographs in dentistry and orthodontics, as they provide a broad view of dental development and enable straightforward assessment of dental maturation (DM). Developing roots and mineralisation stages are clearly visible on DPTs, making DM an accessible indicator during routine examinations (Jeong et al., 2022). Early growth studies, including the work of Hunter (1966), demonstrated that mandibular growth accelerates during adolescence, emphasising the importance of identifying the PGS for optimising treatment timing. This is particularly relevant for Class II malocclusions, where functional appliances are most effective during peak mandibular growth (Baccetti et al., 2006; Franchi et al., 2021; O'Brien et al., 2009). In contrast, in Class III malocclusions, the maxillary sutures begin to close around puberty, which supports the initiation of early treatment before the onset of the PGS (Angelieri et al., 2015; Mandall et al., 2022).

Routine panoramic radiographs often show the developing roots of teeth such as the mandibular canine and first premolar, features that may coincide with periods of active skeletal growth. Dental maturation, however, does not accurately reflect the timing of pubertal growth and should therefore be interpreted as a supplementary rather than definitive indicator for growth assessment. Its value lies in its widespread availability and ease of acquisition during dental visits, although notable individual variation limits its diagnostic precision (Jeong et al., 2022).

Demirjian's method (1973) is one of the most widely applied systems for dental age estimation (Demirjian A, 1973). Developed using French-Canadian children, the method assigns maturity scores to eight stages of tooth calcification, offering a structured and reproducible evaluation system (Figure 2.1) (Adil et al., 2022; Koshy & Tandon, 1998; Nykänen et al., 1998; Nyström et al., 2000). Subsequent studies identified the tendency of the original Demirjian system to overestimate dental age in several populations. This concern led Willems and colleagues to revise and recalibrate the scoring system using Belgian children, resulting in improved accuracy across different ethnic groups (Pandey et al., 2019; Patel et al., 2015; Willems et al., 2001). Dental calcification is considered a relatively stable biological indicator with less susceptibility to environmental and hormonal influences compared with skeletal mineralisation (Demirjian et al., 1985; Jayaraman et al., 2012). Population-specific variations have nevertheless been reported, including among Gujarati and Mumbai populations, where the Willems method demonstrated higher precision than the original Demirjian scores (Figure 2.2).

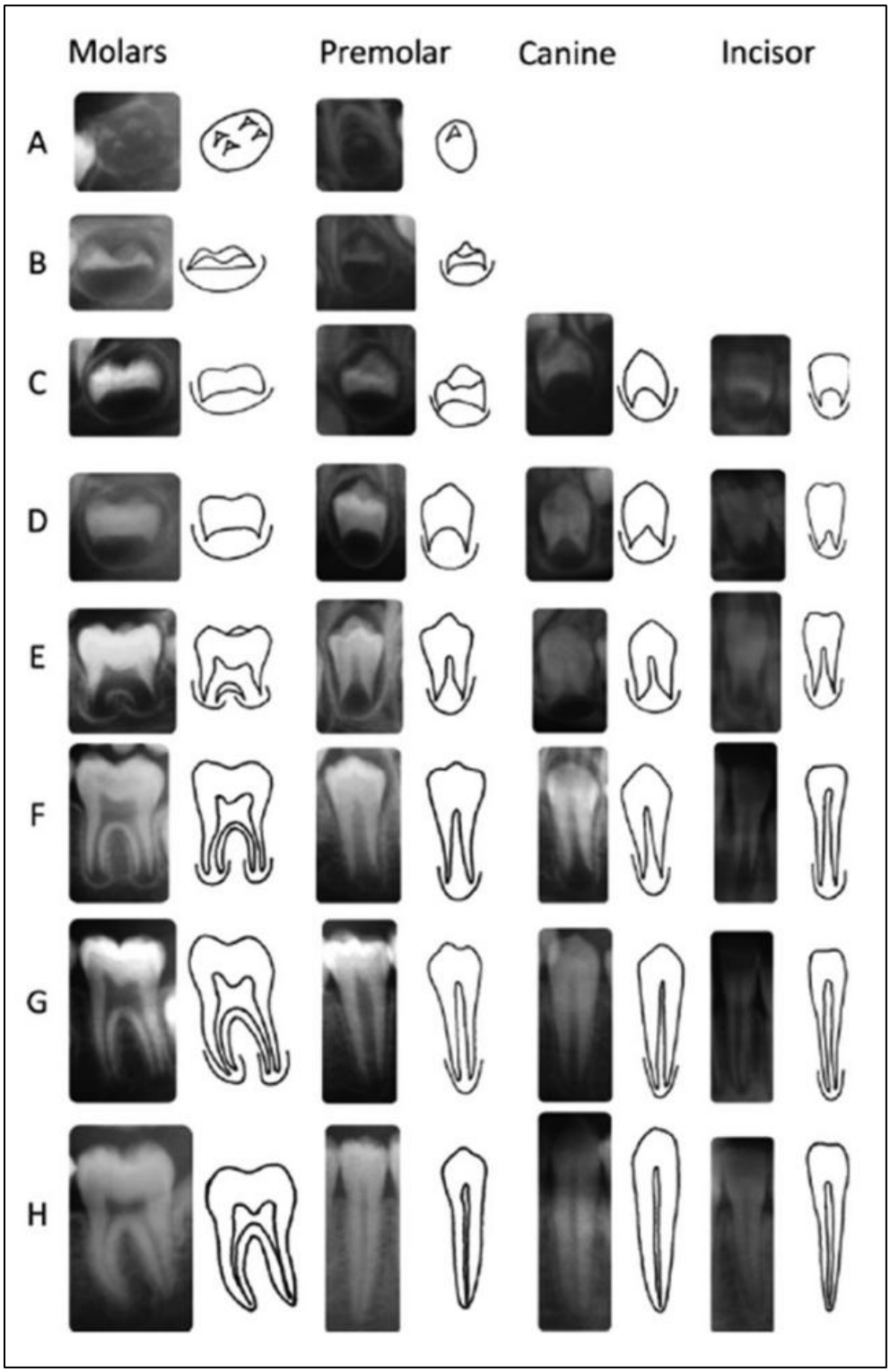


Figure 2.1 Developmental Stages in Demirjian's Method
(Adapted from Demirjian A, 1973)

Stage	Radiographic appearance	Description of stage
Stage 0		Crypt outline visible. No calcification
Stage A		Calcification of single occlusal points without fusion of different calcifications
Stage B		Fusion of mineralization points; the contour of the occlusal surface is recognizable
Stage C		Enamel formation has been completed at the occlusal surface and dentine formation has commenced. The pulp chamber is curved and no pulp horns are visible
Stage D		Crown formation has been completed to the level of the Enamelocemental junction. Root formation has commenced. The pulp horns are beginning to differentiate, but the walls of the pulp chamber remain curved
Stage E		The root length remains shorter than the crown height. The walls of the pulp chamber are straight and the pulp horns have become more differentiated than in the previous stage. In molars, the radicular bifurcation has commenced to calcify
Stage F		The walls of the pulp chamber now form an isosceles triangle and the root length is equal to or greater than the crown height. In molars the bifurcation has developed sufficiently to give the roots a distinct form
Stage G		The walls of the root canal are now parallel, but the apical end is partially open. In molars only the distal root is rated
Stage H		The root apex is completely closed (distal root in molars). The periodontal membrane surrounding the root and apex is uniform in width throughout

Figure 2.2 Schematic Drawings and Brief Description of Stages of Crown and Root Formation used to Score Third Molar Development (Modified Demirjian Method) (Adapted from Willems et al., 2001).

Ethnic origin, climate, and regional characteristics may influence the relationship between dental and skeletal maturation (Uysal et al., 2006). A study conducted in a Southern European population of Spanish origin reported strong correlations between CVM, hand–wrist maturation, and Demirjian's DM, with chronological age also showing strong associations across methods, particularly among females (Sierra, 1987). The mandibular second premolar and second molar were identified as reliable indicators of maturation for males and females respectively, further supporting the usefulness of DPTs for chronological age estimation.

The relationship between dental calcification stages and skeletal maturation indices varies across studies. Some investigations report weak correlations, whereas others describe stronger and clinically meaningful associations (Jayaraman et al., 2012; Sierra, 1987). Gender-related differences have also been reported, with several studies indicating greater consistency of correlations in females than males (Lamparski, 1972; San Román et al., 2002; Uysal et al., 2006). Research in an Iranian population found that dental calcification stages were not reliable for identifying the pubertal growth phase when compared with CVM (Firouzinia et al., 2022). More recent findings suggest that the strongest correlations in females involve the second molars, while in males the mandibular second premolar demonstrates the highest association, although these trends vary across populations and methodologies (Cummaudo et al., 2021).

Dental maturation contributes valuable information for estimating chronological age and serves as a useful adjunctive indicator in clinical settings (Demirjian et al., 1985; Willems et al., 2001). The overall evidence, however, indicates that DM lacks the degree of precision required to identify skeletal maturation or the pubertal growth spurt, as demonstrated by studies reporting inconsistent or population-dependent correlations between dental calcification stages and skeletal maturity indicators such as CVM and hand-wrist method (HWM) (Jayaraman et al., 2012; Sierra, 1987; Uysal et al., 2006). Variability in associations across genders, ethnic groups, and methodological approaches further limits its diagnostic reliability for growth assessment (Cummaudo et al., 2021; Lamparski, 1972; San Román et al., 2002). DM is therefore more appropriately regarded as a supplementary measure rather than a primary tool for determining CVM stage or establishing the optimal timing for growth modification interventions, particularly when more maturation-specific radiographic indicators provide superior accuracy (Baccetti et al., 2005; Firouzinia et al., 2022; McNamara & Franchi, 2018).

2.2.3.2 Hand–Wrist Maturation

Following the assessment of dental maturation on panoramic radiographs, clinicians often turn to more direct skeletal indicators to obtain a clearer understanding of pubertal development. While DPT-based dental maturation provides convenient information obtained from routine dental records, its inability to reliably identify the onset of the pubertal growth spurt has led orthodontists to adopt methods that assess skeletal structures more directly (Demirjian et al., 1985; Firouzinia et al., 2022; Jayaraman et al., 2012). Among these, the HWM has historically been one of the most widely utilised approaches for evaluating skeletal maturation and correlating it with pubertal growth (Fishman, 1982; Houston et al., 1979; Pyle & Greulich, 1959).

The HWM evaluates the sequence of ossification and fusion in the carpals, metacarpals, phalanges and sesamoids, which follow a predictable chronological pattern during growth. These developmental events are commonly assessed using established reference standards, including the Greulich and Pyle atlas, the Tanner–Whitehouse system and pubertal maturation curves described by Bowden and Fishman (Bowden, 1976; Fishman, 1982; Houston et al., 1979; Pyle & Greulich, 1959). Each of these systems differs in purpose and clinical applicability (Figure 2.3). The Greulich and Pyle atlas provides a simple visual comparison for estimating pubertal growth spurt but is limited by ethnic variation and poor sensitivity to pubertal timing (Pyle & Greulich, 1959). The Tanner–Whitehouse system offers a more detailed and quantitative bone-by-bone assessment, yet its complexity reduces practicality in routine orthodontic settings (Houston et al., 1979). Bowden's maturation curves are less widely validated and are used infrequently in modern clinical practice (Bowden, 1976). Fishman's Skeletal Maturation Indicators (SMI), developed specifically for orthodontics, allow identification of the onset, peak and completion of the pubertal growth spurt and are therefore considered the most clinically relevant system for growth modification planning (Fishman, 1982). Nevertheless, the SMI still requires a separate hand–wrist radiograph and remains influenced by observer variability and population-specific differences (Dhiman et al., 2015; Uysal et al., 2006).

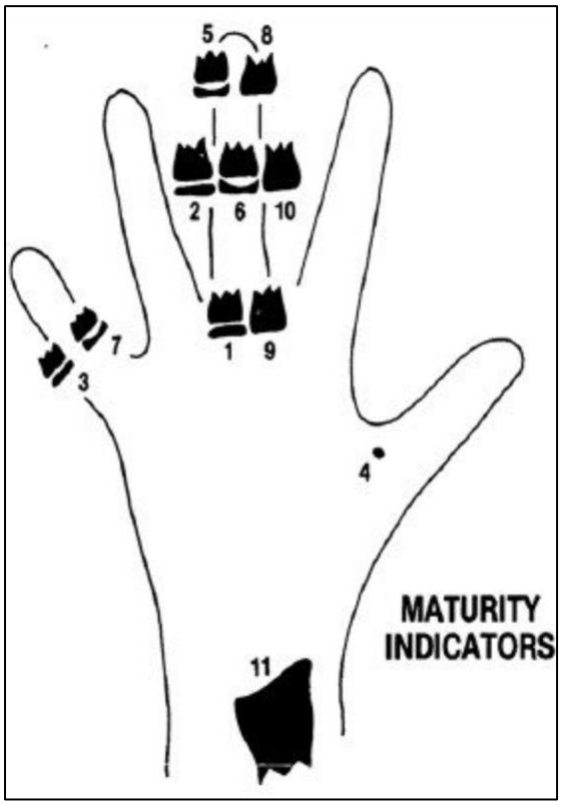


Figure 2.3 Eleven Skeletal Maturity Indicators
of HWM Stages (Adapted from Fishman, 1982)

Despite its longstanding role in assessing skeletal maturation, several limitations reduce the clinical utility of the HWM. Variation in hand–wrist morphology, sexual dimorphism and differences in population growth patterns affect the reliability of CVM stage estimation (Houston, 1980). Furthermore, the skeletal changes observed on hand–wrist radiographs largely correspond to the peak and terminal phases of the pubertal growth spurt, rather than its onset. The inability to detect early pubertal transition limits the method's usefulness for optimally timing orthodontic interceptive or growth-modifying interventions (Houston et al., 1979).

The requirement for a dedicated radiograph represents an additional drawback, as the hand–wrist film is taken solely for maturation assessment and contributes to unnecessary radiation exposure. This conflicts with the principle of maintaining radiation doses as low as reasonably practicable (ALARP), a standard emphasised in dental radiology guidelines (White & Pharoah, 2014). The CVM method is commonly used in orthodontics because it avoids additional radiation exposure and is readily available in routine records.

This integration reduces radiation exposure and has contributed to a shift away from routine use of hand–wrist radiographs in contemporary orthodontic practice (Hassel & Farman, 1995; Valentin, 2007).

Radiographic maturation assessments, including the HWM, also rely heavily on subjective interpretation of bone morphology, sequential ossification stages and shape changes, which introduces both inter-observer and intra-observer variability (Dhiman et al., 2015). These sources of variation, combined with the inability of the HWM to detect the early stages of puberty and its requirement for additional radiation exposure, limit its effectiveness as a primary method for determining CVM stage classification (Hassel & Farman, 1995).

Although the hand–wrist method provides valuable information regarding skeletal development and the timing of the pubertal peak, current evidence indicates that it is not sufficiently precise for identifying the early phases of the pubertal growth spurt or for determining CVM stage classification with the level of accuracy required for growth-based orthodontic decision-making. Its reliance on additional imaging, its limited ability to capture early puberty and its susceptibility to observer variation have contributed to a decline in its use. Contemporary orthodontic practice has increasingly adopted alternative methods such as CVM, which offer greater convenience, reduced radiation exposure and improved applicability for clinical growth assessment (Uysal et al., 2006).

2.2.3.3 Cervical Vertebral Maturation

The limitations of the hand and wrist method prompted the search for a maturation indicator that could be incorporated into routine orthodontic records without the need for additional imaging. Lateral cephalograms are already obtained for virtually all orthodontic patients to evaluate craniofacial morphology, skeletal patterns and dental relationships (Szemraj-Folmer et al., 2021). Their consistent availability created an opportunity to assess skeletal maturity using structures visible on this radiograph alone.

Early investigations into this possibility demonstrated that the morphology of the cervical vertebrae reflects pubertal development and can be used to estimate skeletal maturity (Hassel & Farman, 1995; Lamparski, 1972). Subsequent research refined these observations into structured indices, with the methods developed by Hassel and Farman and the later classification system proposed by Baccetti, Franchi and McNamara

emerging as the most widely implemented in clinical practice (Baccetti et al., 2002, 2005; Ferrillo et al., 2021; Hassel & Farman, 1995). These systems evaluate predictable changes in the shape and contour of the second, third and fourth cervical vertebrae, enabling clinicians to infer growth status directly from a standard cephalometric radiograph.

In orthodontic treatment planning, identifying whether a patient is approaching, experiencing or exiting the pubertal growth spurt is central to optimising the timing of functional appliances, orthopaedic procedures and certain surgical decisions (Franchi et al., 2000). Misalignment between treatment timing and a patient's biological growth pattern can compromise skeletal response and prolong treatment (Hägg & Taranger, 1982). Incorporating CVM analysis into routine diagnostic records therefore provides a convenient and clinically meaningful approach for guiding growth-based interventions.

Evolution of the CVM Method

Systematic evaluation of cervical vertebrae as indicators of skeletal maturity began with the work of Lamparski, who demonstrated that the morphology of the second, third and fourth cervical vertebrae changes in a predictable sequence that corresponds with pubertal development (Lamparski, 1972). This early evidence established cervical vertebral morphology as a potential alternative to the hand and wrist method for assessing growth status.

Building upon Lamparski's findings, Hassel and Farman in 1995 developed a structured maturation index that incorporated two key morphological features: the development of concavities along the inferior vertebral borders and the progressive transformation of vertebral body shape(Hassel & Farman, 1995). This index represented a significant advancement by translating qualitative observations into a practical clinical tool.

Further refinement of cervical vertebral assessment was achieved through the work of Baccetti, Franchi and McNamara, who initially introduced a five stage maturation system and later formalised it into the widely recognised six stage classification (CVM Stages I to VI)(Baccetti et al., 2002). Their updated framework, provided clearer morphological descriptors, improved delineation of pubertal growth phases and enhanced consistency in stage identification (Baccetti et al., 2005; Pratiwi et al., 2025).

The popularity of the modified Baccetti system is largely attributed to its reproducibility and its compatibility with routine orthodontic cephalometric radiographs. The ability to evaluate skeletal maturity directly from existing diagnostic images provides a practical, efficient and radiation free adjunct to growth assessment, contributing to its widespread adoption in contemporary orthodontic practice (Ferrillo et al., 2021; McNamara & Franchi, 2018; Perinetti et al., 2012).

CVM Staging and Morphological Features

The CVM method evaluates skeletal maturity by analysing characteristic morphological changes in the second, third, and fourth cervical vertebrae. These changes primarily involve the progressive development of concavities along the inferior vertebral borders and the transformation of vertebral body shape from trapezoidal to rectangular horizontal, square, and eventually rectangular vertical configurations, reflecting the sequential maturation pattern described by Baccetti *et al.* (Baccetti et al., 2005).

These morphological features correspond to the six stages of the Baccetti CVM classification (Figure 2.4, Table 2.1):

- CVM Stage I: The inferior borders of the C2, C3, and C4 vertebral bodies are flat. The shapes of C3 and C4 are trapezoidal, with the superior border shorter than the inferior border and no concavities. This stage corresponds to the prepubertal growth period, where minimal skeletal maturation changes are evident. Represents a pre-pubertal stage.
- CVM Stage II: A distinct notch begins to appear along the inferior border of C2, whereas the inferior borders of C3 and C4 remain flat. The bodies of C3 and C4 continue to exhibit trapezoidal morphology. This stage is indicative of early changes in vertebral morphology associated with the initiation of pubertal growth.
- CVM Stage III: Inferior border concavities are now visible on both C2 and C3, while the inferior border of C4 remains flat. The shapes of C3 and C4 bodies continue to be trapezoidal. CVM Stage III is typically associated with the acceleration phase of growth.
- CVM Stage IV: All three vertebral bodies (C2–C4) display obvious concavities along their inferior borders. The shapes of C3 and C4 change

from trapezoidal to horizontally elongated rectangles, reflecting a progression towards the deceleration phase of the pubertal growth spurt.

- CVM Stage V: This stage is differentiated from CVM Stage IV by further changes in vertebral body morphology. The shapes of C3 and/or C4 transition to become square, while all three vertebrae continue to exhibit notched inferior borders. This stage is consistent with the maturation phase, where growth deceleration becomes more pronounced.
- CVM Stage VI: At least one of the vertebral bodies, either C3 or C4, develops into a vertically elongated rectangle. In this stage, the length of the posterior border becomes greater than the inferior border. The cortical bone margins also appear more distinctly delineated compared to CVM Stage V, signalling the completion of pubertal growth and the attainment of skeletal maturity.

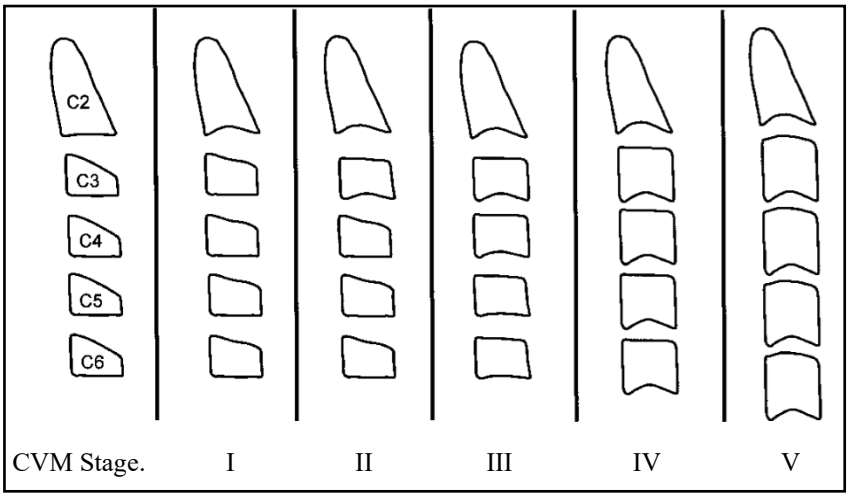
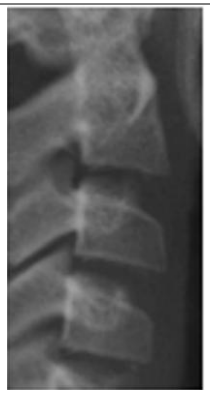
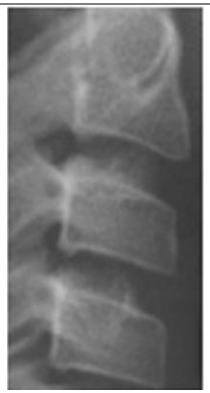
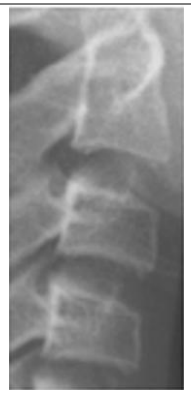
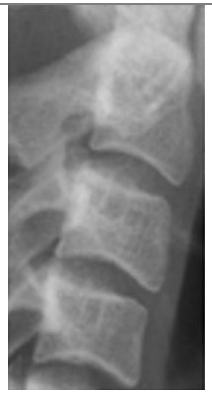
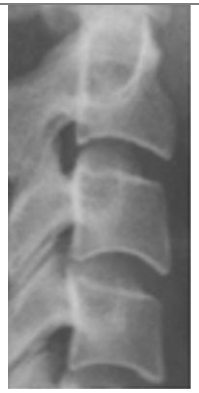
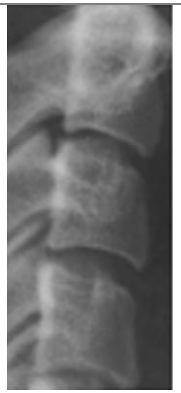


Figure 2.4 Development Stages of Cervical Vertebrae According to Baccetti *et. al.* (Adapted from Baccetti et al., 2002)

Table 2.1
Characteristics of Various Classes of Cervical Maturation Degree.

Stages		CS1	CS2	CS3	CS4	CS5	CS6
Inferior border concavity*	Vertebra						
	C2	Absent	Present	Present	Present	Present	Present
	C3	Absent	Absent	Present	Present	Present	Present
	C4	Absent	Absent	Absent	Present	Present	Present
Body form	C3	Trapezoid	Trapezoid	Trapezoid or Horizontal-rhomboid	Horizontal-rhomboid	Square or Horizontal-rhomboid	Vertical-rhomboid or Square
	C4	Trapezoid	Trapezoid	Trapezoid or Horizontal-rhomboid	Horizontal-rhomboid	Square or Horizontal-rhomboid	Vertical-rhomboid or Square

Research has demonstrated that each CVM stage aligns with a specific chronological age range; however, these age distributions depend strongly on sex and geographic background. Magalhães *et al.* reported consistent global patterns indicating that, except at CVM Stage I, cervical maturation varies by continent (Magalhães et al., 2022). Children in American and European populations reach CVM stages approximately one year earlier than their Asian and African counterparts. Girls also tend to attain CVM stages earlier than boys, reflecting well-established sex-based differences in skeletal maturation (Flores-Mir et al., 2006; Magalhães et al., 2022).

These demographic variations highlight the importance of applying population-specific reference data during orthodontic diagnosis and treatment planning. Ethnic and regional differences in cervical maturation have been shown in several populations. In southern Chinese children, CVM correlates strongly with hand–wrist indicators but weakly with chronological age (Abdulla Alkhal et al., 2008). Indonesian children reach skeletal maturation stages later than Caucasian children, demonstrating the need for culturally and geographically relevant standards (Soegiharto, Moles, et al., 2008). Associations between CVM and dental development have also been observed; for instance, Saudi participants showed significant correlations between CVM stages and tooth calcification levels (Felemban, 2017). These findings reinforce that skeletal maturation patterns vary widely across populations and biological characteristics.

Evidence further suggests that the age at which rapid growth (CVM Stages III–IV) occurs differs across malocclusion types. In a European cohort, rapid growth for Class II patients occurred at an average age of 11.2 years, whereas Class III patients showed a later peak at approximately 12.8 years, indicating that different malocclusion groups may experience distinct growth trajectories (Flieger et al., 2018). For all classifications, growth acceleration typically concluded around age 13.

Together, these findings confirm that the CVM method provides valuable insight into adolescent growth potential; however, its interpretation must be contextualised within sex, ethnicity and diagnostic characteristics to ensure accurate orthodontic treatment planning.

Clinical Application of CVM Staging

The timing of orthodontic and orthopaedic interventions is critically dependent on accurate assessment of the pubertal growth spurt. Different malocclusions exhibit

distinct patterns of skeletal growth responsiveness, making reliable identification of the growth phase essential for optimising treatment outcomes and long-term stability.

For Class II malocclusion**,** skeletal correction is most effective when treatment coincides with the peak in mandibular growth velocity. Functional appliances such as the Twin Block, Herbst, and MARA demonstrate significantly greater skeletal effects particularly mandibular elongation and favourable remodelling of the glenoid fossa when initiated at or immediately before peak height velocity (Franchi et al., 2000, 2013). Initiating Class II therapy too early may result in dentoalveolar rather than skeletal effects, whereas commencing treatment too late reduces skeletal responsiveness and may necessitate camouflage mechanics or surgical correction. CVM staging is therefore valuable for determining whether a patient is in CVM Stages II–III, the interval most closely associated with the ascending and peak phases of pubertal growth, which correspond to optimal timing for Class II functional appliance therapy (Perinetti et al., 2016; Predko-Engel et al., 2015).

For Class III malocclusion, the timing requirements differ substantially. Early orthopaedic intervention typically involving maxillary protraction, face mask therapy, or combination approaches such as RME with protraction is most effective before the pubertal growth spurt, ideally during CVM Stages I–II (Baccetti et al., 2005; Hagg & Pancherz, 1988; Hägg & Taranger, 1982) Class III patients throughout the same time frame display an average age of 12.8 rapid growth. In contrast to patients with Class II, patients with Class III may require a longer period of accelerated growth. Angle estimates that the time difference is approximately 1 year and 6 months. For all types of malocclusions, the growth surge ends around the age of thirteen (Flieger et al., 2018).

Once a patient enters CVM Stage III or beyond, the circummaxillary sutures exhibit reduced responsiveness, diminishing the orthopaedic effects of maxillary protraction. Late presentation or delayed diagnosis may increase the likelihood of deterioration into a more severe skeletal discrepancy, ultimately requiring orthognathic surgery. CVM staging therefore assists clinicians in identifying the window of maximum sutural adaptability and determining whether early intervention remains feasible or whether treatment should shift toward camouflage or surgical planning.

In addition to sagittal discrepancies, CVM assessment also supports the timing of rapid maxillary expansion (RME), as peak skeletal effects are demonstrated when expansion is performed before or during early puberty, when the midpalatal suture

retains its patency, exhibits maximal adaptability and ossification has not yet advanced (Baccetti et al., 2005).

Beyond growth modification, CVM plays an important role in orthognathic surgery planning, particularly for adolescents approaching the end of secondary school. Accurate assessment of skeletal maturity is essential to determine whether facial growth is complete before scheduling definitive surgical correction. Residual mandibular or maxillary growth after surgery is a well-documented risk factor for postoperative instability and relapse (Proffit et al., 2007; Tulloch et al., 2004).

CVM staging provides clinicians with an accessible indicator of whether a patient has reached the post-pubertal or growth-completed phase (CVM Stages V–VI), supporting decisions regarding the appropriate timing for surgery (Baccetti et al., 2005). This becomes particularly valuable when patients and families wish to undergo orthognathic surgery shortly after high school, as CVM assists in determining whether full growth cessation has occurred or whether delaying surgery is advisable (Stahl et al., 2008).

CVM staging enables clinicians to classify a patient as pre-pubertal, pubertal, or post-pubertal, allowing treatment to be tailored to the biological stage rather than chronological age. Appropriate synchronisation of treatment with growth potential can reduce treatment duration, enhance skeletal outcomes, minimise the risk of relapse, and reduce the need for subsequent retreatment or surgical intervention (Perinetti et al., 2012).

Limitations of the CVM Method

The CVM method remains widely used because it is practical, cost effective and easily incorporated into routine orthodontic records without additional radiation exposure (Baccetti et al., 2002; Hassel & Farman, 1995). Despite these advantages, it is constrained by several methodological, biological and technical limitations that affect diagnostic precision, reproducibility and generalisability across clinical settings and populations (Flores-Mir et al., 2006; Nestman et al., 2011; Perinetti et al., 2012). Recognising these limitations highlights the need for improved, standardised frameworks, including artificial intelligence based systems that aim to minimise observer dependent variation and enhance diagnostic consistency (Atici et al., 2022; Gabriel et al., 2009).

Predictive Accuracy for CVM Method

Although CVM stages correlate with the general trajectory of pubertal development, their predictive accuracy for peak height velocity (PHV) and the exact timing of the pubertal growth spurt remains limited. Longitudinal studies indicate that PHV may occur across a wide range of CVM stages, and that identical stages may correspond to different growth velocities among individuals (Franchi et al., 2021; Hagg & Pancherz, 1988). Although CVM correlates with hand–wrist maturation indicators, its precision in predicting PHV is reduced during early adolescence, when vertebral morphological changes remain subtle (Franchi et al., 2000; Hagg & Pancherz, 1988). Consequently, CVM is most clinically reliable between approximately eight and sixteen years of age, when vertebral morphology follows a more predictable pattern of pubertal transformation (Baccetti et al., 2005; Lamparski, 1972). Biological factors such as sex, hormonal variability, nutrition and ethnicity contribute to natural variation in skeletal maturation independent of cervical vertebral morphology (L.-L. Chen et al., 2008; Hagg & Pancherz, 1988; Perinetti et al., 2012). Multiple studies have shown that biological growth timing varies widely among individuals whether assessed through hand–wrist ossification, dental development or vertebral morphology (Demirjian et al., 1985; Fishman, 1982; Perinetti et al., 2012). These physiological differences mean that even accurate CVM staging reflects general maturational trends rather than absolute biological growth potential.

Additional evidence highlights further concerns regarding the accuracy and validity of the CVM method. Low repeatability has been documented, with classification errors occurring frequently even under controlled conditions (Gabriel et al., 2009). Inter-ethnic variation and secular trends in maturation also affect the applicability of CVM staging in diverse populations (Machado et al., 2020). Although several authors support the diagnostic value of cervical vertebral analysis as an alternative to hand–wrist evaluation (Baccetti et al., 2002; Hassel & Farman, 1995), systematic reviews present mixed findings. Three systematic reviews (Cericato et al., 2015; Szemraj-Folmer et al., 2021) reported substantial agreement between CVM and hand–wrist methods, demonstrating that CVM is largely comparable to HWM and may serve as a suitable substitute without requiring additional radiographic exposure. Cericato *et al.* further concluded that the correlation between the two methods is

sufficiently strong that requesting a hand–wrist radiograph solely for skeletal maturation assessment is generally not justified (Cericato et al., 2015).

Despite these encouraging findings, unresolved issues remain regarding the ability of CVM to consistently reflect individual growth patterns and its overall reproducibility, highlighting the need for further high quality research. The reliability, usefulness and validity of the CVM method therefore continue to be debated. Evidence establishing a direct and consistent relationship between vertebral morphology and skeletal maturity is still limited (Chatzigianni & Halazonetis, 2009). A systematic review by Santiago *et al.* emphasised that despite reasonable reproducibility and meaningful correlation with HWM, the available studies represent low level evidence and are insufficient to conclusively validate CVM as a stand-alone indicator of skeletal maturity (Santiago et al., 2012).

i. Ethnic and Population-based Variability

The CVM method was initially standardised using European and North American samples. As research expanded into diverse populations, it became evident that the timing and sequence of cervical vertebral morphological changes vary significantly among different ethnic and geographic groups. Early work by Chen *et al.* demonstrated that Chinese adolescents reached CVM stages later than their Western counterparts, and subsequent analyses confirmed that maturation generally occurs earlier in American and European cohorts than in many Asian and African populations (Y. W. Chen et al., 2020; Magalhães et al., 2022; Soegiharto, Cunningham, et al., 2008). Despite these regional differences, sex related trends remain consistent across studies, with girls achieving CVM stages earlier than boys regardless of ethnicity (Abdulla Alkhal et al., 2008; Soegiharto, Moles, et al., 2008)

More recent investigations reinforce these population specific trends. Multicentre investigations have demonstrated substantial population-specific variation in the chronological age ranges associated with individual CVM stages, with conventional criteria showing reduced validity when applied outside their original reference populations (Kohli et al., 2024; Magalhães et al., 2022). Similar concerns were raised in a 2025 multicentre reproducibility study, which found that examiner agreement decreased in ethnically heterogeneous samples, suggesting that vertebral

morphological features may express differently across populations (Pratiwi et al., 2025).

Population variability also affects the correlation between CVM and other maturity indicators. In southern Chinese samples, CVM demonstrated strong association with hand wrist maturation but weak correlation with chronological age (Abdulla Alkhal et al., 2008). In Indonesian and Saudi populations, researchers observed delayed attainment of peak skeletal maturation compared with European norms, with differences noted across types of malocclusion and between sexes (Felemban, 2017; Soegiharto, Cunningham, et al., 2008). More recently, a 2023 study on Chinese adolescents reported that CVM stages did not consistently align with maturation of other craniofacial structures such as the midpalatal suture, underscoring the complexity of extrapolating growth status from vertebral morphology alone (H. Liu et al., 2023).

Applying a universal CVM classification without population specific calibration risks systematic bias in diagnosing maturation stage and may lead to inappropriate timing of growth modification, rapid maxillary expansion or orthognathic surgery (Machado et al., 2020).

Collectively, the evidence highlights the need for population specific CVM reference data and validation studies. Establishing local norms improves diagnostic accuracy and ensures CVM staging reflects the true growth patterns of the population being treated. Without such contextualisation, CVM should be interpreted with caution, especially in regions characterised by ethnic diversity or mixed ancestry.

ii. Subjectivity and Examiner Variability

One of the most significant limitations of the CVM method is its dependence on subjective visual interpretation. Clinicians are required to evaluate subtle morphological characteristics such as the depth of concavities along the inferior borders of C2, C3 and C4 and the gradual transformation of vertebral body shape. These features change progressively rather than in discrete, sharply defined transitions. As a result, distinguishing between adjacent stages can be difficult when the morphological characteristics are only mildly expressed or when concavities appear asymmetric or shallow. Even trained orthodontists may disagree when assigning stages, as the distinctions between “slight” and “well-defined” concavities or between “trapezoidal”

and “rectangular horizontal” vertebral shapes are open to interpretation (Nestman et al., 2011). This inherent ambiguity makes CVM particularly vulnerable to inconsistent interpretation (Perinetti et al., 2012).

Research consistently shows that the greatest variability occurs at the transitional stages of the CVM system. Stages II and III demonstrate the highest rates of disagreement because concavity formation is emerging but not fully expressed, and vertebral body shapes do not yet match the classic descriptors of the later stages (Nestman et al., 2011). Stage IV also contributes to inconsistency, although to a lesser extent, as clinicians may struggle to determine when the vertebral bodies have transitioned from a rectangular horizontal to rectangular vertical vertebral morphology, which can be difficult to classify which has been shown to reduce inter-observer agreement (Perinetti et al., 2012). Nestman *et al.* reported that nearly 70 percent of inter-observer disagreements occurred at these transitional stages, a finding echoed by Perinetti *et al*., who observed substantial declines in intra-observer reproducibility at the same points. Such discrepancies reflect a fundamental limitation common to biological staging systems, where continuous processes are artificially divided into discrete categorical stages (Flores-Mir et al., 2004; Nestman et al., 2011; Perinetti et al., 2012; Zhao et al., 2012)

Reliability refers to the degree of agreement when a diagnostic tool is applied repeatedly under the same conditions, while reproducibility denotes the consistency of results across different examiners, settings, and contexts (Predko-Engel et al., 2015; Wayne & Ellner, 2010). In the case of CVM, these principles are particularly critical, as clinical decisions regarding treatment timing often hinge on subtle morphological distinctions in the cervical vertebrae.

Reliability, Calibration and Dataset Integrity in CVM Assessment

Accurate assessment of CVM depends not only on the biological relevance of vertebral morphology but also on the consistency with which examiners assign CVM stages. The CVM system categorises skeletal maturation into six ordered diagnostic stages (CVM Stages I–VI), each associated with specific implications for orthodontic treatment timing. For such categorical classifications, reliability metrics that account for agreement beyond chance are essential (Landis & Koch, 1977; McHugh, 2012).

Early reliability studies predominantly employed Cohen's kappa to assess inter- and intra-observer agreement between two examiners. Across studies, Cohen's kappa values have generally ranged from moderate to substantial, although with considerable variation depending on examiner experience, calibration methods and study design. Early reliability studies demonstrated that expert agreement in CVM staging was generally substantial, with inter-examiner kappa values frequently exceeding 0.70, reflecting a shared interpretive framework among calibrated orthodontists. However, subsequent work revealed marked examiner-dependent variability, where agreement ranged from minimal to almost perfect depending on the examiner pair. These findings underscore that CVM staging, despite defined morphological criteria, remains a judgement-based task in which perceptual thresholds and visual inferencing strategies differ between observers (Predko-Engel et al., 2015; Rainey et al., 2016). Multiple studies have also reported that interobserver agreement is consistently lower than intraobserver reproducibility, indicating that clinicians tend to be internally consistent but not necessarily concordant with one another (Gabriel et al., 2009; Perinetti et al., 2012). Intraobserver kappa values frequently approach 0.60 to 0.70, suggesting acceptable repeatability within examiners. This variability has been attributed to ambiguity in CVM morphological descriptors, particularly when differentiating subtle concavity formation and transitional vertebral body shapes such as trapezoidal, rectangular horizontal and square morphologies (Flores-Mir et al., 2004; Nestman et al., 2011).

Validation studies, expert panels and dataset annotation for modelling frequently involve three or more examiners. In contemporary research and dataset development, CVM assessment commonly involves multiple examiners. In such contexts, Fleiss' kappa is preferred over pairwise Cohen's kappa, as it provides a more stable estimate of agreement across three or more raters (Fleiss, 1971; McHugh, 2012). Studies applying Fleiss' kappa have reported moderate to substantial agreement overall, with consistently lower agreement at transitional stages (Hussain et al., 2024; Nestman et al., 2011).

Fleiss' kappa directly evaluates the extent to which multiple observers assign identical CVM stages beyond chance expectation and is now widely recommended for categorical diagnostic methods involving more than two examiners, (Landis & Koch, 1977; McHugh, 2012). Studies applying Fleiss' kappa have consistently demonstrated substantial but imperfect agreement, reinforcing the presence of systematic observer

dependent variability. Nestman *et al.* reported Fleiss' kappa values in the moderate range and observed that most disagreements clustered around transitional stages (Nestman et al., 2011). Another study has. reported an overall Fleiss' kappa of approximately 0.69 for CVM staging, with higher agreement at postpubertal stages where vertebral morphology is more distinct (Sella Tunis et al., 2024). Supporting these findings, a recent systematic review and meta analysis reported pooled interexaminer and intraexaminer kappa values of approximately 0.62 and 0.71 respectively, synthesising results from studies using Cohen's kappa, weighted kappa and Fleiss' kappa, and concluding that CVM demonstrates acceptable reproducibility under controlled conditions despite substantial heterogeneity (Hussain et al., 2024).

By capturing agreement across multiple examiners simultaneously, Fleiss' kappa provides a more realistic estimate of reproducibility than pairwise statistics alone (Predko-Engel et al., 2015; Santiago et al., 2012). As a result, it is increasingly regarded as the preferred reliability metric for multiexaminer CVM studies, large scale dataset annotation and preparation for artificial intelligence training and validation (Ferrillo et al., 2021; Hussain et al., 2024).

Variability in CVM staging has important implications for both clinical decision-making and artificial intelligence development. Inconsistent stage assignment limits the utility of CVM as a reference standard and compromises the integrity of ground truth labels used for AI training (Nestman et al., 2011; Perinetti et al., 2012). Studies further demonstrate that examiner specific interpretation plays a substantial role in stage assignment, with clinicians applying identical reference guidelines differently based on experience and judgement (Jaqueira et al., 2010; Zhao et al., 2012).

Technical imaging factors introduce an additional layer of variability. Differences in radiographic equipment, exposure parameters and patient positioning can affect image quality, while motion artefacts, low contrast and superimposition of anatomical structures may obscure cervical vertebrae and reduce staging accuracy. Metallic orthodontic appliances may further compromise visualisation of key morphological features, particularly in multicentre datasets where imaging protocols are not standardised (Tang et al., 2018).

Consequently, structured examiner calibration, standardised operational definitions and consensus-based annotation have been repeatedly recommended to improve reproducibility and establish more robust reference standards for both clinical use and dataset construction (Flores-Mir et al., 2004; Nestman et al., 2011).

The implications of calibration and consensus are particularly critical for artificial intelligence development. Deep learning systems rely on large volumes of high quality and consistently annotated data, and inconsistent ground truth labels risk propagating human uncertainty into automated models. This often results in acceptable performance for clearly defined stages such as CVM Stage I or VI but reduced reliability for intermediate categories. Inconsistent annotations directly limit model reliability and generalisability.

For artificial intelligence development, these limitations are critical. Deep learning models trained on poorly calibrated or noisy annotations risk reproducing human uncertainty, particularly for intermediate stages, thereby reducing model reliability and generalisability (Esteva et al., 2017). Accordingly, structured examiner calibration and consensus-based annotation are essential for establishing robust CVM reference datasets. Evidence from medical imaging demonstrates that consensus-labelled data substantially improve model performance, reproducibility, and interpretability. These principles are fundamental to the development of clinically reliable AI-based CVM systems capable of supporting standardised growth assessment across diverse populations (Esteva et al., 2017).

Misclassification of skeletal maturity can lead to inappropriate treatment timing, particularly for growth modification therapies where early or delayed intervention may reduce skeletal effectiveness, prolong treatment duration or increase the likelihood of treatment escalation. Improving reproducibility through calibration, consensus and standardisation is therefore essential for research validity, patient safety and clinical outcome optimisation.

In this context, AI-based CVM systems offer a compelling solution. When trained on well calibrated, consensus labelled datasets, automated systems have the potential to reduce examiner dependent variation, improve reproducibility and provide consistent staging across clinical and research environments. Such systems may also increase clinician confidence by functioning as reliable second opinion tools, supporting objective decision making and standardised growth assessment workflows.

Advances and Emerging Technologies in CVM Assessment

Efforts to address the limitations of conventional two dimensional CVM assessment have prompted the development of more sophisticated morphological

techniques designed to quantify vertebral shape more precisely. Researchers have explored statistical shape analysis of cervical vertebrae using both lateral and axial planes of C1 to C4, based on the premise that human growth occurs in three dimensions rather than in a single planar projection (Chatzigianni & Halazonetis, 2009; Yang et al., 2014). Integrating lateral and axial cervical vertebral morphology has the potential to provide a more comprehensive representation of skeletal maturation compared with traditional evaluations that rely solely on qualitative two dimensional interpretation.

Three dimensional assessment, however, requires cone beam computed tomography, which provides the axial and volumetric data necessary for detailed shape analysis. CBCT imaging allows enhanced visualisation of structures that cannot be reliably assessed on lateral cephalograms, such as the odontoid process and the full geometry of the second cervical vertebral body (S. H. Kim et al., 2017). This improved visibility facilitates more precise evaluation of vertebral body proportions, concavity development and other morphological parameters integral to skeletal maturation (Echevarría-Sánchez et al., 2020b). Statistical shape analysis of lateral and axial CBCT derived cervical vertebral forms has demonstrated stronger explanatory power for CVM stage estimation than conventional qualitative staging (Chatzigianni & Halazonetis, 2009; Yang et al., 2014). CBCT derived statistical shape models have demonstrated greater explanatory power for predicting CVM stage classification than traditional qualitative staging systems, with studies proposing quantitative CVM frameworks that outperform earlier methodologies based on two dimensional radiographs (Byun et al., 2015; S. H. Kim et al., 2017).

Despite these potential advantages, several factors limit the clinical feasibility of CBCT based CVM assessment. Foremost among these is radiation exposure. CBCT delivers significantly higher radiation doses than lateral cephalometry, making it unsuitable for routine monitoring of growth in children and contradicting the primary rationale for CVM: avoiding additional imaging by utilising radiographs already obtained for orthodontic diagnosis. Furthermore, current three dimensional studies remain preliminary, often involve restricted sample sizes and lack validation across diverse populations. To date, no CBCT based CVM system has been standardised or adopted into mainstream clinical protocols (Byun et al., 2015). These limitations preclude the integration of three dimensional morphological assessment into routine practice and underscore the continued relevance of optimised two dimensional techniques..

CVM analysis has also expanded in relevance beyond traditional orthodontics. It has been applied in paediatric dentistry, craniofacial research and growth endocrinology to investigate relationships between skeletal maturation, dental development and craniofacial morphology. Perinetti *et al.* for example, demonstrated correlations between CVM stages and dental age indices, supporting its value as a complementary growth marker in multidisciplinary assessment (Perinetti et al., 2013).

2.3 Foundations of Artificial Intelligence and Deep Learning

Artificial intelligence refers to computational systems designed to perform tasks traditionally associated with human cognition, such as pattern recognition, decision making, natural language processing, speech recognition and visual interpretation. Within artificial intelligence, machine learning enables computers to learn from data and improve performance iteratively without explicit programming. With time and additional data, these systems become more effective. Large amounts of digital data are used to build highly accurate machine learning models. Deep learning represents a more advanced subset of machine learning, is more adaptable and employs neural networks with multiple processing layers to extract increasingly abstract and complex features directly from raw data (Rajpurkar et al., 2018; Tang et al., 2018).

These capabilities have led to substantial advances in medical imaging, where deep learning models have demonstrated high diagnostic accuracy and, in some domains, performance comparable to or exceeding that of clinicians, particularly in chest radiograph interpretation and malignancy detection (Jones et al., 2021; Rajpurkar et al., 2018; Wu et al., 2020).

The application of artificial intelligence in medicine has expanded rapidly across multiple disciplines. AI based systems have been used to analyse musculoskeletal imaging, predict survival outcomes, model disease transmission and assist in malignancy detection (García-Pola et al., 2021; Gyftopoulos et al., 2019; Hung et al., 2020; Tsui et al., 2013). In dentistry and orthodontics, artificial intelligence has been increasingly applied in dental imaging, diagnosis, treatment planning and craniofacial anomaly assessment, including cleft lip and palate conditions (Khanagar, Al-Ehaideb, et al., 2021; Schwendicke et al., 2019; Shan et al., 2021). Deep learning algorithms are commonly implemented using neural networks that are conceptually inspired by neural connections seen in the human body.

A neural network consists of interconnected processing units known as perceptrons, each receiving weighted inputs from preceding layers. These inputs are combined and transformed through activation functions to generate outputs that are passed to subsequent layers (Brickley et al., 1998). Typical architectures include an input layer, one or more hidden layers and an output layer. Through successive layers, neural networks learn hierarchical feature representations, allowing them to capture subtle patterns and perform accurate classification across ordered categories (Brickley et al., 1998; Tang et al., 2018).

These characteristics make deep learning particularly well suited to CVM assessment. Neural networks are capable of identifying subtle morphological patterns and distinguishing progressive developmental stages within ordered classification systems. Their hierarchical learning structure aligns naturally with the gradual and continuous nature of vertebral maturation observed on lateral cephalograms, supporting reliable stage classification and reducing dependence on subjective human interpretation.

Several object detection architectures have been explored in medical imaging applications, including Faster R-CNN, Single Shot Detector (SSD), and You Only Look Once (YOLO). Faster R-CNN is a two-stage detector that generally achieves high localisation accuracy through region proposal generation but requires greater computational resources and longer inference times. SSD performs object localisation and classification in a single stage, offering improved computational efficiency and faster processing, although performance may decline when detecting smaller or morphologically subtle anatomical structures (Liu et al., 2021). YOLO-based architectures similarly employ single-stage detection but optimise detection speed and localisation performance through unified feature extraction and prediction pipelines (Redmon et al., 2016). Recent versions such as YOLOv8 demonstrate improved accuracy, efficiency, and adaptability in medical imaging tasks, making them increasingly suitable for real-time anatomical localisation and clinically deployable AI systems (Ueda et al., 2023) . Given the relatively consistent anatomical positioning of cervical vertebrae and the need for efficient localisation within lateral cephalograms, YOLOv8 was selected in the present study as an appropriate balance between detection accuracy, computational efficiency, and scalability.

2.3.1 Limitations of Existing Automated CVM Classification Approaches

Despite growing interest in artificial intelligence for automated CVM classification, current approaches remain constrained by methodological limitations that restrict both their clinical reliability and translational potential. While many studies report promising classification accuracies, typically ranging from 70% to 85%, these performance levels plateau despite increasing model complexity, suggesting that factors beyond algorithm choice are limiting progress (Amasya, Cesur, et al., 2020; Atici et al., 2022; Seo et al., 2021).

A critical and under-addressed limitation in existing CVM automation studies lies in the assumption that training labels represent an unequivocal ground truth. Most deep learning models are trained using single-examiner annotations or small expert panels without formal calibration or reliability assessment. However, CVM staging is inherently subjective, particularly within transitional stages where morphological features overlap and biological change occurs gradually rather than discretely (Nestman et al., 2011; Perinetti et al., 2012). As a result, inter- and intra-observer variability introduces label noise that is propagated directly into supervised learning models. In such contexts, improvements in model architecture alone cannot overcome inconsistencies embedded within the reference annotations.

This limitation is particularly consequential for transitional stages, notably CVM Stage II–III and CVM Stage III–IV, which have consistently demonstrated the lowest levels of examiner agreement in manual assessments. Several automated studies report reduced sensitivity and precision at these stages, mirroring the same diagnostic uncertainty observed among human raters (Amasya, Yildirim, et al., 2020; Atici et al., 2022). These findings indicate that model misclassifications at transitional boundaries are not solely technical failures but reflect unresolved ambiguity within the underlying ground truth. Without addressing annotation reliability, automated systems risk learning examiner bias rather than true biological patterns of skeletal maturation.

In addition, many existing AI-based CVM studies adopt end-to-end convolutional neural network architectures that operate on full lateral cephalometric images. While such approaches may achieve high training accuracy, they often demonstrate limited generalisability due to background anatomical noise and reliance on spurious features unrelated to vertebral morphology. Structures such as the mandible, cranial base, dentition, and soft tissues may inadvertently influence classification,

particularly in imbalanced datasets where certain CVM stages dominate. The absence of explicit anatomical localisation further reduces interpretability and limits clinical trust in model outputs.

Importantly, relatively few studies integrate reliability science into the development of automated CVM systems. Formal examiner calibration, agreement analysis, and consensus-based annotation protocols remain uncommon in the literature, despite being well established in clinical research as prerequisites for valid diagnostic assessment. Without such frameworks, reported model performance metrics may overestimate true diagnostic capability and fail to reflect real-world clinical variability.

Collectively, these limitations highlight a critical gap in the current literature: the need for automated CVM classification systems that are grounded in reliable, expert-calibrated annotations and that explicitly address the anatomical and biological sources of diagnostic ambiguity. Addressing this gap requires a methodological shift away from purely algorithm-driven optimisation toward clinically informed system design, integrating multi-expert consensus, structured calibration, and anatomically constrained analysis pipelines. The present study was designed to address these limitations directly by embedding reliability assessment and expert calibration as foundational components of automated CVM model development.

2.3.2 Dataset Imbalance, Oversampling Strategies, and Model Generalisability in Automated CVM Classification

A persistent challenge in automated CVM classification lies in the inherent imbalance of available datasets. In clinical orthodontic populations, early developmental stages (CVM Stage I and II) are underrepresented because lateral cephalometric radiographs are rarely acquired in very young children unless clinically indicated. Conversely, later stages are overrepresented due to the age distribution of patients typically seeking orthodontic treatment. This skewed class distribution introduces systematic bias during supervised machine learning, where models tend to favour majority classes and demonstrate reduced sensitivity for minority and transitional stages.

Class imbalance is a well-recognised problem in medical imaging AI, where high overall accuracy may mask poor performance in clinically critical but underrepresented categories (Chawla et al., 2002; He & Garcia, 2009). In CVM

classification, this issue is particularly consequential because transitional stages coincide with the pubertal growth spurt, where diagnostic precision directly influences treatment timing. Several AI-based CVM studies have reported reduced classification performance at these stages, even when overall accuracy appears acceptable, highlighting the limitations of training on imbalanced datasets (Amasya, Yildirim, et al., 2020; Mathew et al., 2022; Seo et al., 2021).

To mitigate imbalance, resampling techniques such as Random Oversampling (ROS) and the Synthetic Minority Over-sampling Technique (SMOTE) have been widely applied in medical AI. ROS increases the representation of minority classes by duplicating existing samples, while SMOTE generates synthetic samples through interpolation (Chawla et al., 2002). Although these approaches can improve sensitivity for minority classes, they also introduce risks, including overfitting and limited feature diversity, particularly when applied to small or noisy datasets (Buda et al., 2018). In medical imaging tasks involving subtle anatomical variation, oversampling may stabilise class distribution without fundamentally improving discriminatory capacity.

Beyond data-level solutions, architectural considerations play a critical role in addressing imbalance-related performance degradation. End-to-end deep convolutional neural networks trained directly on full radiographs are prone to learning spurious correlations from background anatomy rather than clinically relevant structures, especially when minority classes are scarce (N. Kazimierczak et al., 2024; Polizzi & Leonardi, 2024; Schwendicke et al., 2019). Recent studies have therefore advocated for anatomically guided pipelines that explicitly localise regions of interest prior to classification, reducing background noise and improving generalisability (S. Li et al., 2022; Z. Li & Wang, 2025; Seo et al., 2021).

Generalisation remains a central concern in automated CVM classification. Models trained on imbalanced or institution-specific datasets often exhibit performance degradation when applied to external populations, reflecting sensitivity to demographic, ethnic, and imaging-protocol variability (Schwendicke et al., 2019; Topol, 2019). These limitations underscore the importance of combining dataset-level strategies with robust model design, expert-validated ground truth, and biologically informed feature representations to achieve clinically reliable performance.

Collectively, existing literature indicates that dataset imbalance is not merely a statistical inconvenience but a fundamental barrier to reliable automated CVM staging. Addressing this challenge requires a balanced integration of data curation strategies,

architectural design, and validation frameworks, rather than reliance on resampling techniques alone. These considerations form the methodological basis for subsequent model development and evaluation strategies explored in later chapters.

2.4 Bibliometric Analysis of Artificial Intelligence in Orthodontics

2.4.1 Growth of AI Research in Orthodontics

Recent years have witnessed a rapid expansion of artificial intelligence applications across orthodontics, driven by advances in digital imaging, computational power, and the availability of large clinical datasets (Khanagar, Al-ehaideb, et al., 2021; Schwendicke et al., 2019). While narrative and systematic reviews have provided valuable insights into specific AI models and clinical applications, they are limited in their ability to objectively map the structure, evolution, and intellectual landscape of a rapidly developing research field (Gowri Priya & Thirugnanam, 2024; M. K. Lee et al., 2024).

Bibliometric analysis offers a complementary, quantitative approach by systematically analysing publication outputs, citation networks, and keyword co-occurrence patterns, thereby identifying research trends, influential authors and institutions, and emerging thematic clusters (Pirri et al., 2020). In orthodontics, bibliometric methods have increasingly been applied to evaluate the growth and diversification of AI-related research, particularly in areas such as diagnostic imaging, cephalometric landmark detection, malocclusion classification, and treatment outcome prediction (Jiang et al., 2025; Schwendicke et al., 2019).

Despite this expansion, the degree to which artificial intelligence has been applied to growth assessment and skeletal maturity evaluation remains unclear. Given the fundamental importance of growth timing in orthodontic diagnosis and treatment planning, this uncertainty warrants systematic investigation. Bibliometric evaluation provides an objective means of determining whether growth-related applications, including CVM assessment, have been proportionately explored relative to their clinical significance.

Accordingly, this bibliometric analysis aims to characterise temporal trends in AI orthodontic research, identify dominant methodological and application-based

themes, and evaluate the relative positioning of growth and skeletal maturation within the broader AI orthodontic literature. By mapping these patterns, this section establishes contextual justification for the present study and highlights gaps that motivate the development of a robust, standardised AI-based CVM classification system.

2.4.2 Global and Temporal Publication Trends

The trajectory of research in artificial intelligence applications within orthodontics has expanded substantially over the past decade. Early work between 2000 and 2018 was characterised by slow and methodologically narrow output, largely centred on traditional machine learning approaches such as support vector machines (SVMs), texture analysis, and engineered morphological features. These methods were limited by constrained computational capacity, relatively small institutional datasets, and the absence of large-scale annotated imaging corpora. Following the adoption of deep learning, particularly CNNs after 2018, publication volume increased sharply, mirroring broader developments in medical imaging and digital dentistry. Peaks between 2021 and 2023 coincide with the maturation of imaging architectures, open-source frameworks, and cross-disciplinary collaborations between dental schools and engineering faculties (Figure 2.5). This sustained rise reflects both technological momentum and the recognition of orthodontics as a clinically suitable domain for AI-driven diagnostics and decision support.

Geographically, publication activity is unevenly distributed, with concentration in East Asia, North America, and parts of the Middle East (Figure 2.6). The United States, China and Turkey represent leading contributors, supported by strong academic–engineering partnerships and large clinical imaging datasets (S. Li et al., 2022). Asian centres, including India, Saudi Arabia, and South Korea, have produced methodologically rigorous studies emphasising skeletal maturity assessment, cephalometric analysis, and digital workflow integration. Emerging contributions from Germany and United Arab Emirates highlight broader international participation, often through region-specific datasets addressing demographic variability. This geographic imbalance suggests that research capacity, annotation expertise, and computational infrastructure remain clustered in well-resourced institutions.

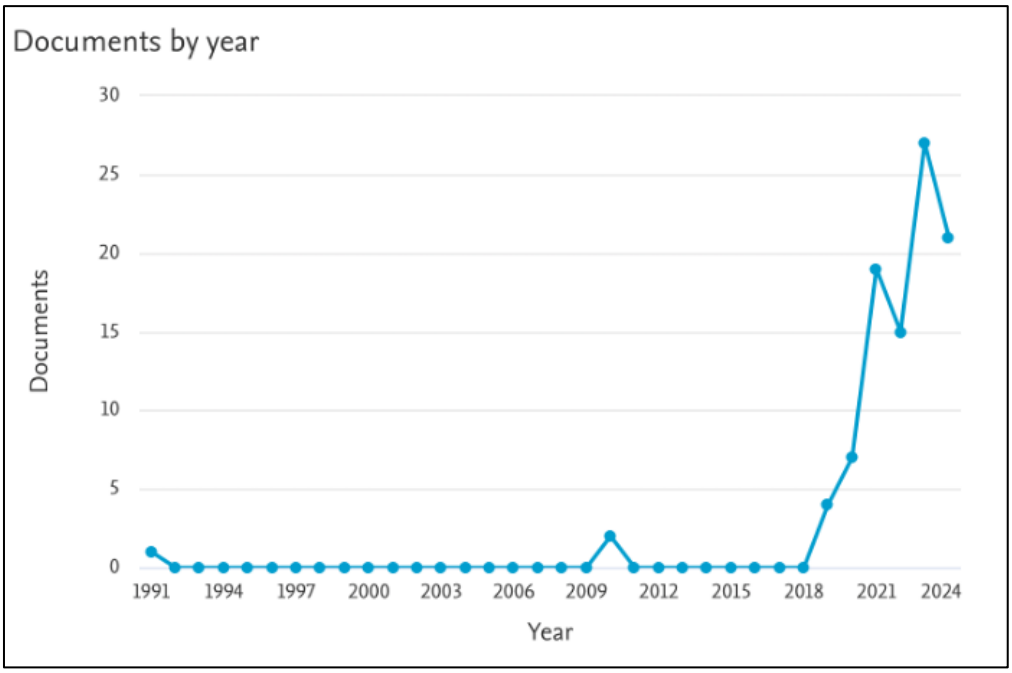


Figure 2.5 Annual Publication Trends in AI and Orthodontics (2000–2024).

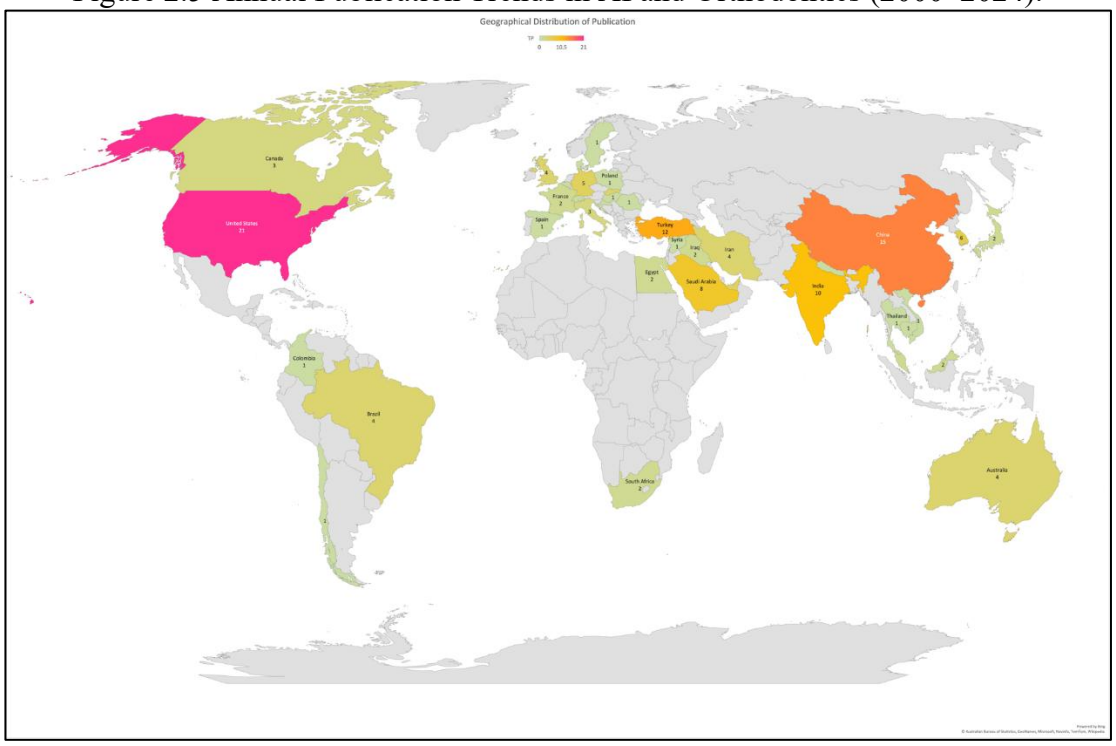


Figure 2.6 Global Distribution of AI in Orthodontics Research.

At the institutional level, universities with strong dental schools and established biomedical engineering collaborations, such as Sichuan University and the University of Illinois at Chicago, consistently ranked among the leading contributors. Notably, collaborative networks frequently extended beyond dental faculties to include computer science and biomedical engineering departments, indicating that AI in orthodontics is

advancing through interdisciplinary mechanisms rather than isolated clinical research units (Figure 2.7). This institutional pattern reinforces the broader observation that methodological progress is being driven by engineering–dentistry partnerships, where imaging expertise, computational infrastructure, and annotation capacity are co-located (Norman, Rosli, et al., 2025).

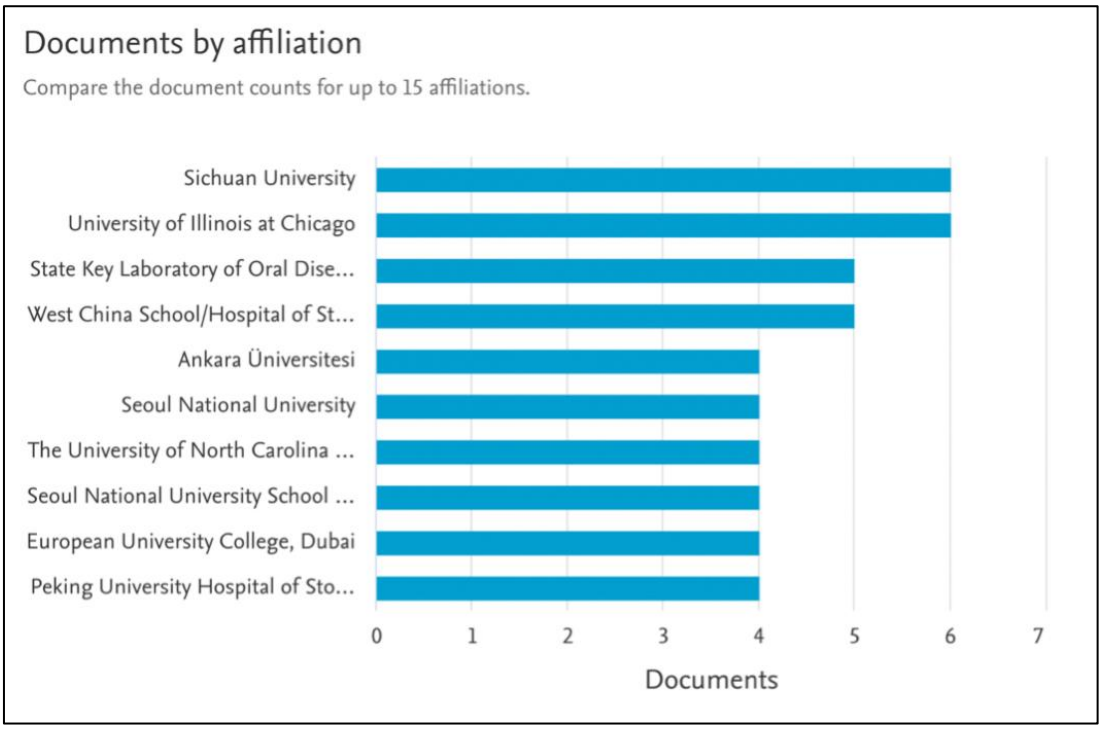


Figure 2.7 Documents by Institutional Affiliation.

2.4.3 Key Journals, Authors, and Citation Impact

Journal-level analysis shows that orthodontic AI research is predominantly disseminated through specialty orthodontic journals, indicating that the field remains disciplinarily anchored. However, emerging contributions from multidisciplinary venues suggest increasing methodological input from adjacent biomedical and computational domains. Orthodontics and Craniofacial Research, American Journal of Orthodontics and Dentofacial Orthopedics, Seminar in Orthodontics, and Angle Orthodontist emerged as the most active publication venues, reflecting a disciplinary consolidation of the field. Notably, Progress in Orthodontics demonstrated the strongest citation indicators (CiteScore, SJR, SNIP, Q1) within the dataset, suggesting that high-impact contributions remain anchored within orthodontic specialty journals rather than diffusing primarily into interdisciplinary or computational journals.

This disciplinary anchoring suggests that AI uptake in orthodontics has been driven by clinically orientated research priorities rather than purely methodological experimentation. However, authorship patterns reveal increasing participation from interdisciplinary teams involving dentistry, biomedical imaging, and computer science, which may explain the shift from handcrafted feature-based models to CNN-based and hybrid architectures in recent years. Citation impact remains modest relative to mature AI fields such as radiology and oncology, indicating that while orthodontic AI is expanding and consolidating intellectually, it has not yet achieved full field-level maturity or broad international influence (Table 2.2).

Table 2.2
The Ten (10) Most Influential Active Journals within the Domain.

Source titles, publisher (number)	Cite score 2023	SJR 2023	SNIP 2023	JCI	CQ	IF
Orthodontics and Craniofacial Research, Wiley (10)	5.3	0.97	1.68	0.92	Q2	2.4
American Journal of Orthodontics and Dentofacial Orthopedics, Mosby-	4.8	1.28	1.566	1.22	Q1	2.7
Elsevier (7)	NA	0.41	NA	NA	NA	NA
Seminar in Orthodontics, WB Saunders (7)	6.4	1.45	1.998	1.5	Q1	3.0
Angle Orthodontist, Research Foundation (4)	0.8	0.174	0.261	0.20	Q4	0.5
APOS Trends in Orthodontics, Scientific Scholar (4)	3.5	0.685	1.300	1.03	Q1	2.6
Korean Journal of Orthodontics, Korean Association of Orthodontists (4)	7.3	1.392	1.962	0.67	Q1	3.5
Progress in Orthodontics, Springer (4)	5.3	0.508	0.924	0.56	Q2	2.5
Applied Sciences (Switzerland), MDPI (3)	4.7	0.667	0.912	0.87	Q1	3.0
Diagnostics, MDPI (3)	3.5	0.606	0.784	0.95	Q2	2.4
Healthcare Switzerland, MDPI (3)						

Notes: SJR = SCImago Journal Rank, SNIP = Source Normalised Impact per Paper, JCI = Journal Citation Indicator, CQ = Category Quartile, IF = Impact Factor, NA= Not Available, MDPI = Multidisciplinary Digital Publishing Institute

Citation impact analysis further identified the most influential articles within the domain (Table 2.3). The majority of highly cited works originated from imaging automation tasks, including cephalometric landmark detection, CBCT-based analysis, and skeletal maturity assessment. Treatment-related tasks, such as extraction decision-making and orthognathic outcome prediction, formed a secondary cluster of influential publications, while systematic and scoping reviews contributed to field consolidation and conceptual organisation.

This distribution mirrors the thematic structure identified in the keyword co-occurrence analysis, suggesting that diagnostic imaging has served as the methodological entry point for AI adoption in orthodontics, with clinical decision-making applications emerging downstream. Collectively, these patterns reinforce the interpretation that imaging con stitutes the most mature domain of orthodontic AI, whereas clinical translation and treatment planning remain comparatively nascent. Notably, two of the most cited articles involved CVM-based skeletal maturity assessment, underscoring the clinical salience of growth-related diagnostics and supporting the relevance of automated CVM assessment within AI-driven orthodontics.

Table 2.3
Top 10 Cited Articles in AI in Orthodontics (Minimum 40 Citations).

Authors	Journal Title	Publication	Cites/ author	Cites/ year	Type
Kunz et al.	Artificial intelligence in orthodontics: evaluation of fully automated cephalometric analysis using customized convolutional neural network	J Orofac Orthop	138	34.5	Article
Xie et al.[13]	Artificial neural network modeling for deciding if extractions are necessary prior to orthodontic treatment	Angle Orthod	119	8.5	Article
Kök et al.[47]	Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics	Prog Orthod	94	18.8	Article
Khanagar et al.[48]	Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - a systematic review	J Dent Sci	87	29	Review
Li et al.[20]	Orthodontic treatment planning based on artificial neural networks	Sci Rep	86	17.2	Article
Bichu et al.[42]	Applications of artificial intelligence and machine learning in orthodontics: a scoping review	Prog Orthod	79	26.3	Review
Amasya et al.[28]	Cervical vertebral maturation assessment on lateral cephalometric radiographs using artificial intelligence: comparison of machine learning classifier models	Dentomaxillofac Radiol	56	14	Article

Mohammad-Rahimi et al.[49]	Machine learning and orthodontics, current trends and the future opportunities: a scoping review	Am J Orthod Dentofacial Orthop	50	16.7	Review
Chen et al.[35]	Machine learning in orthodontics: introducing a 3s auto-segmentation and auto-landmark finder of CBCT images to assess maxillary constriction in unilateral impacted canine patients	Angle Orthod	48	12	Article
Tanikawa and Yamashiro[64]	Developmental of novel artificial intelligence systems to predict facial morphology after orthognathic surgery and orthodontic treatment in Japanese patients	Sci Rep	47	15.7	Article
Suhail et al.[61]	Machine learning for the diagnosis of orthodontic extractions: A computational analysis using ensemble learning	Bioeng.	47	11.75	Article

2.4.4 Research Clusters and Thematic Structure

Keyword co-occurrence mapping (Figure 2.8) revealed three dominant thematic clusters. The red cluster combines diagnostic imaging, cephalometric analysis, and skeletal maturation tasks, reflecting close methodological and conceptual links between landmark detection, growth assessment, and age estimation. The blue cluster represents algorithmic and performance-oriented terminology centred around machine learning models, prediction, and diagnostic accuracy. The green cluster corresponds to clinically contextualised applications, including treatment planning, orthognathic surgery, dental procedures, and workflow-related terminology, indicating emerging interest in embedding AI systems within real-world orthodontic care pathways.

These clusters represent intersecting yet asymmetrical lines of research development. Imaging-related tasks within the red cluster have undergone substantial methodological refinement, transitioning from handcrafted features to deep learning architectures with clinically competitive performance. In contrast, clinically embedded applications within the green cluster appear more heterogeneous and downstream, reflecting the need to integrate diagnostic outputs with treatment decision-making processes. Temporal patterns derived from publication trends further indicate a shift from technically constrained image analysis toward more integrated clinical applications, although deployment in routine orthodontic workflows remains limited.

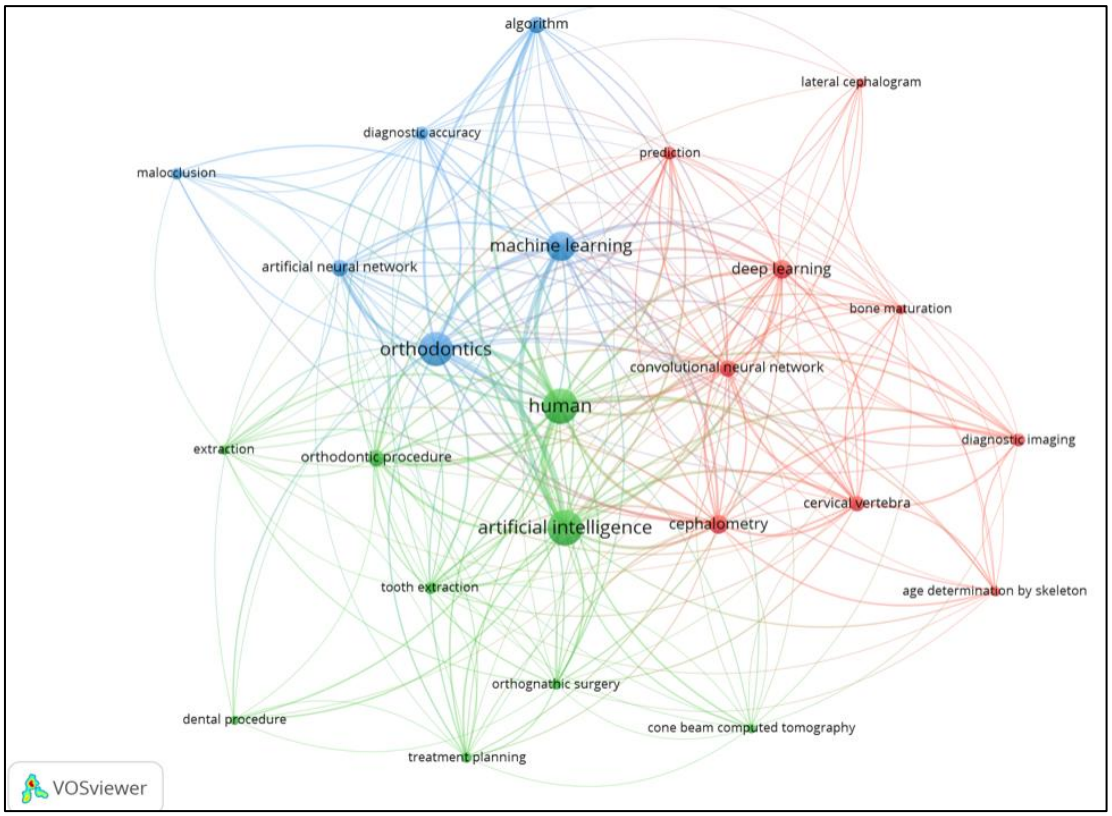


Figure 2.8 Keyword Co-occurrence Network.

2.4.5 Positioning of Growth and Skeletal Maturation within AI Orthodontics

Within the broader bibliometric landscape, growth and skeletal maturation occupy a clinically central yet methodologically uneven position. Applications such as CVM and hand–wrist analysis have sustained scholarly attention due to their direct influence on treatment timing and orthopaedic intervention decisions (Baccetti et al., 2005; Schwendicke et al., 2019). These tasks seek to reduce the subjectivity of traditional staging methods and to standardise orthodontic decision-making, positioning growth assessment as a foundational target for AI-driven diagnostics.

However, growth-related AI research remains fragmented. In contrast to cephalometric landmark detection, growth assessment displays lower standardisation in dataset construction, annotation protocols, validation strategies, and reliability reporting. Many studies operate within isolated institutional contexts, limiting scalability and external validity. Bibliometric patterns therefore suggest that skeletal maturity assessment exerts disproportionate conceptual influence yet remains underdeveloped in terms of reproducibility, methodological maturity, and deployment readiness.

2.4.6 Methodological and Translational Gaps

The bibliometric findings highlight several limitations constraining clinical translation in orthodontic AI. First, most studies rely on single-centre datasets that are demographically homogeneous and institutionally specific, raising concerns about generalisability. Second, annotation quality and reliability are inconsistently reported, despite well-documented inter- and intra-observer variability in skeletal staging tasks. Without calibrated annotation or consensus-driven ground truth, label noise undermines training efficacy and confidence in performance metrics, particularly at transitional maturation stages. Third, while retrospective performance evaluations are common, prospective validation and workflow integration remain rare, restricting the alignment of AI systems with real-world diagnostic pathways. Explainability is also underdeveloped, despite its relevance to clinician trust and medico-legal defensibility in growth-related decision-making.

2.4.7 Implications for This Thesis

The bibliometric analysis provides several insights that inform the rationale and design of the present study. First, growth and skeletal maturation assessment represent high-impact yet methodologically immature domains within orthodontic AI, suggesting a clear opportunity for innovation. Second, persistent challenges related to annotation reliability, dataset diversity, and multi-centre validation justify the methodological emphasis on calibrated expert annotation and external evaluation adopted in this thesis. Third, the observed disconnect between retrospective model performance and clinical workflow integration underscores the need for systems that align with practical diagnostic contexts. Finally, the relative underdevelopment of AI-based CVM assessment provides strong justification for investigating a standardised, explainable classification system tailored to orthodontic growth evaluation.

These findings situate the development of an AI-based CVM classification system within a broader research landscape and provide a natural transition to Section 2.5, which examines state-of-the-art systems and technical considerations relevant to automated skeletal maturation analysis.

2.5 Artificial Intelligence and Automation in CVM Assessment

2.5.1 Artificial Intelligence in Orthodontics

Orthodontics is inherently data-driven, with diagnosis, growth assessment, and treatment monitoring anchored in image-based evaluation and temporal pattern recognition. These characteristics render the specialty uniquely amenable to computational approaches capable of modelling complex visual morphology, estimating growth-related change, and supporting decision making under uncertainty.

While early applications of AI centred on cephalometric landmark detection and semi-automated analysis, subsequent developments have broadened the scope to include skeletal maturation assessment, treatment planning, aligner design, and automated monitoring of tooth movement and treatment progress (J. M. Lee et al., 2024; Schwendicke et al., 2019). This section reviews the development of AI in CVM assessments, from early machine learning approaches to contemporary deep learning systems, with particular emphasis on CVM staging.

2.5.2 Early Applications of Artificial Intelligence in Orthodontics

Initial applications of artificial intelligence in orthodontics emerged in the early 2000s and were largely centred on automating repetitive diagnostic tasks. Cephalometric landmark identification was one of the earliest targets, given its fundamental role in orthodontic diagnosis and treatment planning. Early systems relied on rule based methods, statistical modelling, and classical machine learning techniques that used hand engineered features and template matching to locate anatomical landmarks on lateral cephalograms (Kunz et al., 2020). While these approaches reduced the burden of manual tracing, their performance was highly sensitive to image quality, magnification, and patient positioning.

Subsequent studies introduced support vector machines (SVMs), k nearest neighbour classifiers, and artificial neural networks (ANNs) to improve accuracy and robustness. These methods demonstrated incremental gains in landmark detection and skeletal classification, but their reliance on predefined feature extraction limited scalability and adaptability across datasets (Caldas et al., 2010). Moreover, most early investigations were constrained by small sample sizes, typically fewer than 200

radiographs, reflecting ethical considerations related to radiographic exposure and limited access to digitised datasets. Despite these constraints, early work established a conceptual foundation for automated orthodontic diagnostics and demonstrated that computational systems could replicate certain aspects of clinical reasoning.

SVMs, k-nearest neighbour classifiers, and ANNs were subsequently employed to improve diagnostic accuracy. These methods demonstrated modest gains in identifying cephalometric points and classifying skeletal relationships, but their reliance on pre-defined feature extraction limited scalability (Caldas et al., 2010). Moreover, most studies were restricted to small datasets of fewer than 200 radiographs, a reflection of the challenges of medical data acquisition and ethical concerns over repeated radiographic exposure. Despite these limitations, early work provided a conceptual foundation for the automation of orthodontic diagnostics and demonstrated that computers could replicate aspects of clinical decision-making.

Alongside diagnostic applications, early machine learning models were also explored for treatment planning tasks. Bayesian networks and decision trees were tested on small retrospective datasets, often achieving prediction accuracies in the range of 70–80% (J. H. Lee et al., 2020). Although these figures were insufficient for clinical deployment, they highlighted the potential of AI to complement human expertise in treatment planning, particularly in cases with borderline indications.

2.5.3 Evolution of Automated CVM Classification

The evolution of automated CVM classification reflects broader advances in artificial intelligence within dentistry and orthodontics. Early machine learning applications focused on structured image based diagnostic tasks, particularly cephalometric analysis. These early approaches relied on manually engineered geometric or texture-based features extracted from lateral cephalograms and applied conventional statistical classifiers for diagnostic support. Early machine learning systems established feasibility but were constrained by limited datasets, sensitivity to image quality, and dependence on handcrafted features, restricting clinical scalability and generalisability.

Deep learning approaches demonstrated superior performance in orthodontic imaging tasks by learning hierarchical morphological features directly from radiographic data, overcoming the limitations of manually engineered descriptors

(Esteva et al., 2017; Hung et al., 2020). This capability is particularly suited to orthodontic imaging, where anatomical changes are subtle, continuous, and difficult to capture using predefined descriptors.

Earlier automated CVM classification systems commonly utilised conventional machine learning algorithms such as Support Vector Machines (SVMs), which rely on manually engineered features extracted from cervical vertebral morphology. SVMs are advantageous when datasets are limited, as they can perform well with smaller sample sizes and lower computational requirements. However, their performance depends heavily on feature selection and manual landmark identification, which may introduce operator variability (Khanagar, et al., 2021). In contrast, Artificial Neural Networks (ANNs), particularly deep convolutional neural networks (CNNs), enable hierarchical feature learning directly from imaging data without explicit manual feature engineering. ANN-based systems demonstrate improved adaptability and scalability in complex imaging tasks but generally require larger annotated datasets and greater computational resources (Atici et al., 2023). The transition from SVM-based to ANN-based architectures therefore reflects the broader evolution of medical imaging AI toward automated feature learning and end-to-end classification frameworks.

Cephalometric landmark identification was one of the earliest orthodontic applications to benefit from deep learning. Convolutional neural networks achieved localisation accuracy comparable to experienced clinicians while substantially reducing analysis time, effectively transforming cephalometric analysis into a largely automated workflow (S. Chen et al., 2020). Building on these advances, object detection architectures such as Faster R-CNN, YOLO, and RetinaNet enabled simultaneous localisation and classification of orthodontically relevant anatomical structures under variable imaging conditions (H. Li et al., 2022; J. Liu et al., 2021).

Parallel progress was observed in skeletal maturity assessment using hand wrist radiographs. Early machine learning approaches applied texture analysis and support vector machines to estimate CVM stage classification (Caldas et al., 2010). More recent deep learning models achieved high accuracy in classifying pubertal growth phases, in some cases approaching expert consensus (Amasya, Yildirim, et al., 2020). These findings demonstrated that skeletal maturation could be inferred from radiographic morphology using automated methods, providing a methodological foundation for subsequent application to CVM assessment.

Despite these advances, automation of CVM assessment progressed more slowly. Early CVM models were predominantly semi automated systems designed to replicate conventional staging through manually derived geometric or morphological features. These models were typically trained and validated against a small number of human observers, limiting robustness and external validity. Performance remained constrained by reliance on handcrafted features and subjective reference labels.

Subsequent neural network based CVM systems sought to overcome these limitations by learning discriminative morphological patterns directly from imaging data. A systematic review by Mathew *et al.* reported that neural network based CVM classification systems can achieve diagnostic accuracy comparable to human observers when benchmarked against expert defined reference standards (Mathew et al., 2022). However, substantial variability in reported performance persists, driven primarily by differences in dataset size, annotation consistency, and observer agreement (Mathew et al., 2022).

Recent work has moved beyond direct image classification toward clinically inspired two-stage pipelines. Amasya *et al.* reported CNN-based CVM classification accuracy exceeding 80%, although misclassification remained common at transitional stages (Amasya, Yildirim, et al., 2020). Atici *et al.* improved generalisability using larger datasets and transfer learning, but ambiguity persisted at biologically overlapping intermediate stages (Atici et al., 2022). To better reflect clinical reasoning, object detection frameworks such as YOLO have been introduced to localise C2–C4 vertebrae prior to staging, reducing interference from surrounding structures and improving robustness (Kazimierczak et al., 2024; Mathew et al., 2022; Schwendicke et al., 2019). Hybrid pipelines combining localisation with secondary classification have demonstrated improved performance compared with single-stage approaches (Seo et al., 2021).

Ensemble strategies have also been explored to improve robustness. Studies have demonstrated that combining multiple prediction streams, including convolutional models and morphology based features, enhanced accuracy and interpretability, albeit at increased computational cost (Mathew et al., 2022; Mohammad-Rahimi et al., 2021). Table 2.4 summarises key developments in artificial intelligence based CVM classification reported in the literature.

Within the broader context of dentistry, artificial intelligence has been widely applied to diagnosis and treatment planning for caries, periodontal disease, and oral

cancer, automated dental image analysis including tooth segmentation and orthodontic assessment, computer aided prosthetic design, dental materials analysis, and large scale dental informatics to support clinical decision making (Brickley et al., 1998; Hung et al., 2020; Khanagar, Al-Ehaideb, et al., 2021; Makaremi et al., 2019; Shan et al., 2021). In comparison, automated CVM assessment remains comparatively underdeveloped, reflecting both the complexity of vertebral morphology and the specialised expertise required for reliable staging.

Despite substantial methodological progress, existing AI-based approaches to CVM staging have not yet achieved consistent clinical readiness. Limitations related to dataset imbalance, annotation reliability, transitional ambiguity, and generalisability persist across studies, indicating that further refinement is required before automated systems can be reliably deployed in orthodontic practice. These developments highlight the methodological shift toward clinically aligned, multi-stage AI pipelines for skeletal maturity assessment.

A synthesis of published CVM studies reflects a distinct four-phase developmental trajectory associated with increasing automation, data availability, and clinical alignment (Table 2.4):

Phase 1: Semi-Automated ANN Era (2019–2020)

Early CVM models adopted semi-automated neural network classifiers trained on manually cropped vertebral regions and handcrafted morphological features. Performance was benchmarked against one or two observers, often with apparent accuracies exceeding 90%, but these figures lacked external validity due to limited dataset diversity, categorical staging assumptions, and heavy dependence on manual preprocessing. Conceptually, these models corresponded to the classical machine learning paradigm prevalent in cephalometric analysis, adapted to a more biologically continuous maturation process.

Phase 2: Hybrid Landmark + CNN Era (2021–2022)

The transition to convolutional neural networks introduced localisation and landmarking pipelines that reflected clinical interrogation of vertebral morphology. Landmark detectors, region-of-interest extractors, and secondary classifiers formed

hybrid multi-stage workflows that improved performance and interpretability relative to earlier ANN systems. However, models continued to be constrained by single-observer annotations, class imbalance, and categorical staging frameworks that struggle to represent transitional CVM stages where biological continuity blurs categorical boundaries.

Phase 3: Fully Automated End-to-End Era (2023–2025)

Subsequent work adopted end-to-end architectures that removed manual preprocessing, incorporated transfer learning, expanded training cohorts, and evaluated performance using multi-metric clinical indicators such as ICC, weighted κ, and AUC. Collapsing CVM into fewer maturity categories improved staging coherence and reflected emerging recognition that conventional six-stage CVM schemes may not represent an optimal computational problem formulation. Misclassification during transitional intervals persisted, reflecting shared limitations between algorithmic inference and expert annotation.

Phase 4: Multi-Observer, QCVM, Gender-Aware, Multi-Input Era (2025→)

Emerging systems incorporate multi-observer consensus to improve annotation reliability, quasi-continuous CVM representations (QCVM) to model maturation as a biological continuum, and demographic modifiers such as sex and age to reflect known growth patterns. Attention mechanisms, multimodal inputs, and explainability tools signal a shift toward clinical decision-support rather than categorical classification. This phase corresponds to broader trends in medical imaging AI, where interpretability, multi-factor inference, and clinical integration supersede model-centric accuracy alone.

Table 2.4
Summary of AI Advances in CVM Staging from 2019-2023.

Year	Study	Dataset & Observers (Obsv)	AI Method	Performance Accuracy	Limitations	Key Insight / Contribution
2019	Makaremi (FR) (Makaremi et al., 2019)	n=1870 Obsv: 1	Semi-auto; Entropy filter	Internal: 90%	Single obsv Cropped images	Image filtering improved discrimination
	Kök (TR) (Kök et al., 2019)	n=419 Age: 8–17y Obsv: 1	Semi-auto; ANN (7 nets); Manual landmarks: 32	Internal: 94%	Hassel–Farman method; Manual landmarking	Age & gender improved staging
2020	Amasya (TR) (Amasya, Yildirim, et al., 2020)	n=647 Age: 10-30y Radio: 3 Ortho: 1	Semi-auto; ANN; Manual landmarks: 54	Internal: 87%; External:58%	Observer variability Moderate accuracy external validation	ANN comparable to human observer (50-62%)
2021	D.W. Kim (KR) (D. W. Kim et al., 2021)	n=920 Age: 6-18yo Specialists: 2	Semi-auto; CNN + ROI + segmentation Manual landmarks: 26	Internal: 62.5%	Manual ROI Low accuracy internal validation	Multi-stage pipeline
	Seo (KR) (Seo et al., 2021)	n=600 Age: 6-19y Radio= 1	Semi-auto; CNN	Internal: 94%	Limited dataset Data augmentation ↑ generalisation	↑ Dataset sample
2021	Zhou (CN) (Zhou et al., 2021)	n=1080 Obsv: 1	Fully-auto; DetNet + ResNet50	Internal: 71% ICC: 98%	Overfitting; Single observer	Landmark-driven end-to-end
2022	Mohammad-Rahimi (IR) (Mohammad-Rahimi et al., 2022)	n=890 Ortho: 2	Fully-auto; Oversampling	Internal: 6-Class: 62% 3-Class: 83% External: 51%	Class imbalance Low accuracy external validation	End-to-end framework ↓ Data augmentation in minority groups

2022	Atici (US) (Atici et al., 2022)	n=1018 OMFS: 1 Ortho: 1	Fully-auto; 3 nets	Internal: 85%	Reduced 6→5 classes No external validation	Public dataset (CF Growth Centre)
2022	Liao (CN) . (Liao et al., 2022)	n=900 Age: 7-25y Radio/ortho: 3	Fully-auto; iCVM	Internal: 69–84%	High-res; Observer variability	Dataset size > model complexity
2022	S. Li (CN) (S. Li et al., 2022)	Age: 5–18y	Fully-auto; 4 nets	Internal: 67%	Small dataset	Heatmaps for landmark localisation
2022	Agarwal (IN) **(Agarwal & Agarwal, 2022)**	n=383 Age: 10-36y	Fully-auto; ResNet50/VGG19/custom	Internal: Custom: 94% ResNet50: 91% VGG19: 89%	Data leakage risk Heavy data augmentation	Overfitting due to small dataset Feature extraction dominates
2023	Khazaei (IR) (Khazaei et al., 2023)	n=1846 Age: 5-18y Ortho: 2	Fully-auto	Internal: 3-Class: 82% 2-Class: 93% External: 51%	Data imbalance Low accuracy external validation	End-to-end comparable ↓ Data augmentation
2023	Atici (US) (Atici et al., 2023)	n=1012 Age: 4-29y Obsv: 1 Ortho: 1	Fully-auto; 4 nets	Internal: Males: 75% Female: 82%	↓ 6 Classes to 5 Directional filters to CV edges + age No external validation	Gender specific
2023	Akay (TR) (Akay et al., 2023)	n=588 Age: 8-22y Radio: 2	Fully-auto; custom DCNN using Keras, Tensorflow	Internal: 58%	Moderate accuracy No external validation	Custom frameworks lag big CNNs
2023	H. Li (CN) (H. Li et al., 2023)	n=10200	Fully-auto; 3-stage pipeline	External: 70% AUC 0.94	Complexity CVM Stage II,III: 60-63%	Optimised multi-module

				κ 0.65, κw 0.84		
2024	Shoari (Shoari et al., 2024)	n=663	Fully-auto; ResNet18, custom	Internal: ResNet: 87.5% Custom: 81%	Small sample Public dataset AAOF growth centre	Growth pattern conditioning
2024	Mohammed (Mohammed et al., 2024)	n=1200 Age: 8-16y	Fully-auto	Internal: 97%	Lack 3D	Suggests 3D integration
2025	Zhang (CN) (Zhang et al., 2025)	n=1732 Dentist: 4	Fully-auto; CDSNet	Internal: 90%	Definitional ambiguity Used pneumonia data	Refine network to detect object and mitigate noise
2025	Jiang (CN) (Jiang et al., 2025)	n=2100	Fully-auto; CvNet + QCVM Landmarks: 19	Internal: 69.5% Landmark error 0.66mm	Limited modalities	Landmark QCVM Multi modal - HW xray,age
2025	Abdulqader (Abdulqader et al., 2025)	n=3600	Fully-auto; YOLOv3+ResNet101 Landmarks: 13	Internal: QCVM: 78%	Clinical link pending	Towards clinical applicability
2025	Ramnarayan. (BK et al., 2025)	n=525 Age: 7-17y	Fully-auto; VGG CNN Landmarks: 39	Internal: 85%	Small sample Heavy data augmentation	↓ Data augmentation ↑ Dataset sample

2.5.4 Key Challenges in AI-Based CVM Classification

Despite methodological advances, several challenges continue to limit the clinical readiness of artificial intelligence–based CVM classification. One of the most persistent constraints is class imbalance. Early stages of skeletal development, particularly CVM Stage I and II, are consistently underrepresented in available datasets due to ethical and clinical restrictions on radiographic imaging of young children. Strategies such as random oversampling and synthetic data augmentation, including SMOTE, have been applied to mitigate class imbalance, but these techniques do not fully compensate for the scarcity of diverse and representative clinical data and may fail to improve generalisability to unseen populations (Buda et al., 2018; Chawla et al., 2002; Khazaei et al., 2023).

Transitional stage ambiguity presents an additional and clinically important limitation. CVM stages II and IV correspond to biologically continuous transitions in vertebral morphology, where radiographic features overlap and categorical boundaries are inherently indistinct. This ambiguity contributes to reduced inter and intra observer agreement among clinicians and remains a significant source of error in automated systems trained on human annotated labels (Nestman et al., 2011; Perinetti et al., 2012). Although examiner calibration and the inclusion of adjunctive maturity indicators have improved reproducibility in manual assessment (Flores-Mir et al., 2006), artificial intelligence models continue to struggle with these transitional categories, reflecting shared limitations between human and algorithmic interpretation.

Generalisation and external validation further constrain the deployment of AI based CVM systems. Most published models have been trained and evaluated using single institution datasets characterised by homogenous imaging protocols and population demographics. Skeletal maturation patterns vary according to sex, ethnicity, and environmental factors, raising concerns that models trained on restricted datasets may encode population specific biases and show reduced performance when applied in different clinical contexts (Y. W. Chen et al., 2020; Wiens et al., 2019). Without robust multi centre validation, the external validity of current CVM AI systems remains uncertain (Topol, 2019).

Explainability and clinician trust represent additional barriers to adoption. Deep learning models are often perceived as opaque, producing predictions without transparent justification, which limits clinician confidence in high impact decisions

such as orthodontic treatment timing (Shortliffe & Sepúlveda, 2018). Explainable artificial intelligence techniques, including saliency mapping and Gradient weighted Class Activation Mapping, have been proposed to enhance interpretability by visualising image regions contributing to model predictions (Selvaraju et al., 2017). Evidence from medical imaging literature indicates that explainable outputs improve clinician oversight and acceptance, yet application of these approaches in orthodontics remains limited (Tonekaboni et al., 2019).

Finally, workflow integration remains a critical challenge. For AI systems to progress beyond experimental prototypes, they must integrate seamlessly into routine clinical environments, providing real time decision support, intuitive visualisation of outputs, and opportunities for clinician interaction and override. Studies in radiology and dermatology demonstrate that hybrid clinician AI interfaces enhance efficiency and trust while preserving professional judgment (Esteva et al., 2017; Topol, 2019). Comparable implementation frameworks tailored to orthodontics are still largely absent, limiting the translational impact of current CVM AI models.

These challenges demonstrate that technical accuracy alone is insufficient to ensure meaningful clinical adoption of AI-based CVM systems. Beyond performance metrics, AI tools intended for orthodontic decision support must also satisfy broader digital health standards that emphasise reproducibility, transparency, interpretability, and trustworthiness.

2.6 Translational Barriers and Requirements for Clinical-Grade CVM AI Systems

The development of automated CVM assessment systems requires consideration of the biological, methodological, and operational factors that influence both annotation quality and clinical utility. Although traditional CVM assessment provides a clinically relevant and radiation-free method for estimating skeletal maturation, its translational limitations become more pronounced when reframed as a computational task involving ordinal classification and localised morphological cues. These translational domains are analytically distinct yet interdependent; biological ordinality shapes annotation quality, annotation quality influences model generalisation, and generalisation interacts with interpretability and decision-support requirements in clinical deployment. Contemporary medical AI literature emphasises that successful clinical deployment

requires alignment across these domains rather than optimisation of predictive accuracy alone (Topol, 2019). The following sections review key translational barriers and requirements for clinical-grade CVM AI systems, including biological and annotation constraints, generalisation and external validation, interpretability and workflow integration, and decision-support and equity considerations. Together, these domains highlight the gap between algorithmic performance and clinically meaningful deployment, providing the conceptual foundation for a calibrated, consensus-driven, localisation-based approach to automated CVM staging.

2.6.1 Biological Complexity and Annotation Constraints

Manual CVM assessment relies on the clinician's ability to interpret subtle morphological changes in the C2–C4 vertebrae, where features evolve continuously rather than discretely across developmental stages (Gabriel et al., 2009; Nestman et al., 2011; Perinetti et al., 2012). These changes are particularly challenging to discern within transitional intervals, where morphological cues do not align neatly with categorical staging criteria. Consequently, transitional stages represent a primary source of variability and disagreement among clinicians, even when standardised criteria are applied (Chatzigianni & Halazonetis, 2009; Flores-Mir et al., 2006). Although examiner training and calibration improve reliability to some extent, subjectivity remains inherent to visual interpretation and influences both clinical decision-making and research validity.

However, contemporary evidence increasingly recognises that the clinical utility of CVM staging should not be overstated. Variability in vertebral morphology, transitional overlap between stages, and inconsistent inter-observer reproducibility continue to limit the precision of CVM as a standalone growth predictor (Kazimierczak et al., 2024). Accordingly, CVM assessment is best interpreted as one component of a multifactorial diagnostic approach rather than a definitive determinant of treatment outcome or growth potential.

From an AI development perspective, these biological and interpretive characteristics translate into annotation constraints that affect dataset construction and ground truth formation. Weakly calibrated annotations introduce label noise that propagates into model training and reduces generalisability (Esteva et al., 2017). This challenge is not unique to CVM; similar issues have been observed in other medical

imaging tasks where ordinal or continuous biological processes are discretised for classification purposes. For CVM specifically, annotation quality is influenced not only by examiner expertise but also by the interpretive strategies clinicians employ when weighing competing morphological cues within transitional stages. Such heterogeneity may lead to inconsistencies in dataset labels, particularly when annotations are derived from single-centre or single-expert settings.

The inherently ordinal nature of CVM staging further complicates conventional classification approaches that assume nominal categories. Ordinality implies that misclassifications between adjacent stages (e.g., CVM Stage II→III) differ in clinical significance from misclassifications across distant stages (e.g., CVM Stage II→V). However, standard cross-entropy-based training objectives treat all errors as equivalent, potentially misaligning optimisation with clinical priorities. These considerations support the use of annotation strategies and modelling approaches that account for biological ordinality, morphological subtleties, and the interpretive behaviour of domain experts.

2.6.2 Generalisation, Bias, and External Validation

Generalisation remains a central concern in the development of automated systems for CVM assessment. Across medical imaging domains, models trained on single-centre datasets have demonstrated performance degradation when evaluated on independent cohorts, reflecting sensitivity to demographic and institutional variation (Y. W. Chen et al., 2020; Wiens et al., 2019). Similar concerns arise within orthodontics and craniofacial research, where skeletal maturation and craniofacial growth patterns differ according to sex, ethnicity, and environmental context. This issue is particularly salient in multi-ethnic clinical contexts such as Malaysia, where Malay, Chinese, Indian, and minority ethnic populations contribute to craniofacial diversity and skeletal variation (Department of Statistics Malaysia, 2023; Khazaei et al., 2023). Previous CVM research has predominantly focused on Western cohorts, limiting claims of global applicability and raising questions regarding the representativeness of training samples (Flores-Mir et al., 2004).

In addition to demographic factors, imaging heterogeneity has been identified as a determinant of model performance in dental and medical AI. Variations in exposure parameters, patient positioning, magnification, and soft tissue superimposition can alter

vertebral representation, thereby affecting both human and automated interpretations (Kelly et al., 2019). Recent work in medical imaging suggests that systems trained under narrow imaging conditions may perform suboptimally in routine clinical environments, where cephalograms exhibit natural variability in quality and presentation (Esteva et al., 2021).

External validation has been increasingly emphasised as a mechanism for assessing generalisation and identifying dataset-specific biases prior to deployment. Within digital health, independent evaluation across sites or populations has been proposed as a means of supporting claims of robustness, fairness, and translational relevance (Dzau & Ginsburg, 2016; Topol, 2019). Observations from dermatology, radiology, and pathology suggest that generalisation issues are neither unique to CVM nor limited to dental applications, with performance degradation commonly reported when models are transferred across institutions or populations (Schwendicke et al., 2019; Topol, 2019). This body of literature highlights the importance of multi-centre dataset construction, demographic and imaging heterogeneity, and external validation when considering the translational trajectory of AI systems for clinical use.

2.6.3 Interpretability, Transparency, and Clinician Trust

Interpretability has been identified as a key consideration for clinical adoption of AI in dentistry and medical imaging, where decision-making processes must remain transparent and defensible to clinicians and patients. Studies in radiology, dermatology, and ophthalmology have outlined concerns regarding the opacity of deep learning models and the difficulty of reconciling such opacity with clinical accountability frameworks (Esteva et al., 2021; Topol, 2019). Within orthodontics, clinicians routinely justify staging and treatment decisions based on identifiable morphological cues, suggesting that AI systems intended to support clinical decision-making should reflect similar interpretive strategies. This expectation has motivated the incorporation of explainability techniques, such as Class Activation Mapping and Grad-CAM, in dental AI research to visualise anatomical regions contributing to model predictions and thereby facilitate clinician understanding (Schwendicke et al., 2019).

Workflow integration represents an additional dimension in translational orthodontic AI. Unlike offline benchmarking environments, clinical workflows require models to operate alongside diagnostic reasoning, patient communication, and

treatment planning. The orthodontic management of growth modification, for example, hinges on accurately identifying pubertal windows of skeletal maturation that influence the timing and modality of intervention. Systems that merely predict maturational stages without regard to workflow demands may therefore provide limited utility to clinicians. Digital health literature has highlighted the importance of designing AI tools that align with existing clinical practices and augment, rather than disrupt, clinician reasoning (Dzau & Ginsburg, 2016; Topol, 2019).

Trust emerges as a unifying theme across interpretability and workflow considerations. For clinicians to rely on automated CVM staging, predictions must be not only accurate but also explainable and compatible with orthodontic diagnostic processes. The convergence of transparency, workflow compatibility, and interpretive alignment reflects broader expectations for medical AI, wherein systems are evaluated not solely on performance metrics but on their capacity to integrate into complex clinical ecosystems. Within orthodontics, where diagnostic imaging, growth assessment, appliance selection, and treatment timing interact in longitudinal care pathways, such integration requires careful consideration of how AI-derived outputs are communicated and interpreted within the clinic.

2.6.4 Workflow Integration and Clinical Utility

The literature increasingly frames medical AI as a mechanism for decision support rather than automation. In orthodontics, decisions related to skeletal maturation influence treatment timing, modality, and expected outcomes. These decisions are inherently clinical and require contextual judgement. Digital health scholarship has argued that AI systems should complement clinician expertise by standardising elements of diagnostic assessment, reducing variability, and enhancing reproducibility while preserving clinician agency and responsibility (Dzau & Ginsburg, 2016; Topol, 2019). Within the context of CVM, decision support may take the form of consistent staging, improved timing estimation, or standardised assessment across practitioners and institutions.

Equity considerations have also emerged as a relevant theme in medical AI. Generalisation challenges arising from demographic and imaging variability have the potential to produce disparate model performance across subgroups, particularly in multi-ethnic clinical settings. In orthodontics, where skeletal maturation trajectories

vary across populations and environments, equitable access to decision-support tools necessitates attention to demographic diversity in dataset construction and validation. The literature suggests that AI systems trained on homogeneous cohorts may inadvertently perpetuate disparities, whereas heterogeneous datasets and multi-institutional collaborations may support more equitable distribution of benefits (Schwendicke et al., 2019; Topol, 2019).

Reproducibility and standardisation represent additional domains relevant to decision support and equity. Manual CVM staging has historically been influenced by examiner experience and interpretive strategies, contributing to interobserver variability. AI-driven standardisation may reduce variability and support more consistent treatment decision-making across clinicians. Such standardisation aligns with broader digital health goals of improving reproducibility, supporting guideline adherence, and reducing unwarranted variation in care. These themes collectively situate automated CVM assessment within the broader trajectory of AI-enabled decision support in healthcare.

2.7 Rationale for AI-based CVM Assessment

The domains reviewed in Sections 2.6.1 to 2.6.4 illustrate the multifaceted challenges associated with translating automated CVM assessment into clinical practice. Biological ordinality, annotation variability, demographic and imaging heterogeneity, interpretability expectations, workflow integration, and equity considerations collectively shape both the design and evaluation of clinically viable CVM systems. These domains provide the contextual foundation for the present rationale, which articulates the methodological and conceptual requirements underpinning AI-based CVM development.

From an artificial intelligence perspective, these challenges underscore the importance of reliable ground truth. AI models trained on weakly calibrated or inconsistent annotations risk inheriting human ambiguity, resulting in reduced reproducibility and limited generalisability (Esteva et al., 2017). Addressing this requires large, well-curated datasets developed through structured calibration and consensus-based annotation, ideally incorporating population diversity to mitigate demographic bias and improve external validity (Nestman et al., 2011; Perinetti et al., 2012).

Given the biological continuity of skeletal maturation, clinically meaningful CVM assessment must also accommodate intrinsic uncertainty, particularly at transitional stages. Rather than enforcing rigid deterministic classifications, AI systems intended for clinical support should incorporate transparency and uncertainty awareness. Explainable AI techniques that visualise anatomically relevant regions contributing to predictions facilitate clinician oversight and align with principles of reliability science and trustworthy diagnostic reasoning (Selvaraju et al., 2017; Tonekaboni et al., 2019).

Artificial intelligence therefore represents a logical advancement for CVM assessment by enabling standardised, reproducible, and scalable interpretation of routine lateral cephalograms without additional radiation exposure. Although existing studies report promising performance, limitations related to dataset heterogeneity, annotation reliability, and insufficient external validation continue to restrict clinical deployment (Atici et al., 2022; Khazaei et al., 2023; Mathew et al., 2022).

Accordingly, the present study aims to develop a fully automated CVM classification system trained on a large, expertly calibrated dataset of digital lateral cephalograms and evaluated against multi-expert consensus and conventional CVM assessment. By integrating structured calibration, consensus-based annotation, and contemporary deep learning architectures, this work seeks to advance CVM assessment from an inherently subjective radiographic method toward a reliable, clinically aligned decision-support framework.

This rationale directly informs the methodological design presented in Chapter 3, where dataset construction, annotation protocols, model development, validation strategy, and performance evaluation criteria are specified in alignment with these translational considerations.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Study Design and Ethical Approval

This study employed an observational, cross-sectional, and retrospective analytical design to develop and evaluate an automated CVM classification system using deep learning. The model is designed to replicate expert CVM staging from a single time-point radiograph and does not attempt to infer longitudinal growth velocity or duration. The lateral cephalometric radiographs were collected from the Diagnostic Imaging Unit at the following venues:

Faculty of Dentistry in the following institutions:

1. Universiti Teknologi MARA (UiTM)
2. Universiti Malaya (UM)
3. Universiti Sains Malaysia (USM)
4. Universiti Kebangsaan Malaysia (UKM)
5. Universiti Sains Islam Malaysia (USIM)

Private dental clinics across Malaysia:

6. Klinik Pergigian Presidental
7. The Dental Atelier

Ethical approval was obtained by the UiTM Research Ethics Committee referenced REC/06/2023 (PG/MR/205 dated 15th June 2023), which waived the requirement for informed consent due to the retrospective nature of the study. Data management and analysis were conducted in accordance with the principles outlined in the ICH Good Clinical Practice Guidelines, Malaysian Good Clinical Practice Guidelines, and the Declaration of Helsinki. This study was conducted in accordance with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) 2024 update to enhance transparency, reproducibility, and reporting quality in medical AI research (Tejani et al., 2024b). At no point during or after data collection did the authors have access to information that could identify individual participants. Ethnicity (Malay,

Chinese, and Indian) was extracted from patient records and included as a demographic variable for descriptive characterisation of the study sample.

Table 3.1 outlines the overall research design, mapping the research objectives to the corresponding study phases, research activities, and expected outcomes. Figure 3.1 summarises the overall research framework underpinning the development and validation of the proposed CVMaXX system, structured into three sequential phases. Phase 1 (Preliminary Study) focuses on establishing a robust and reliable ground truth through multi-expert calibration, assessment of inter- and intra-observer reliability, and refinement of CVM visual assessment criteria. Phase 2 (Model Development) addresses the design and implementation of a fully automated two-stage deep learning architecture, incorporating vertebral localisation, morphological feature extraction, and CVM stage classification. Phase 3 (Implementation and Validation) evaluates the clinical performance of the automated system on independent datasets, comparing its accuracy and agreement against calibrated orthodontic expert assessments. Collectively, this phased approach ensures methodological rigor, traceability between research questions and objectives, and progressive validation from expert consensus to automated clinical application.

Table 3.1
Mapping of Research Objectives to Study Phases, Activities, and Outcomes.

Research objectives	Phase	Activities	Outcome
RO1: To evaluate inter- and intra-observer reliability in CVM classification among multiple orthodontic experts	Phase 1 Preliminary Study	Independent CVM staging by calibrated orthodontic experts; intra- and inter-observer agreement analysis using Cohen's kappa and Fleiss' kappa	Quantified reliability metrics identifying stages with highest observer variability
RO2: To establish a calibrated, consensus-based ground truth dataset for CVM staging	Phase 1 Preliminary Study	Expert calibration sessions; consensus adjudication for discrepant cases; dataset refinement	High-quality, consensus-labelled CVM dataset suitable for AI model training

RO3: To develop a fully automated CVM classification system using a two-stage deep learning architecture	Phase 2 Development	Stage 1 cervical vertebra localisation using object detection; Stage 2 CVM classification using deep learning and gradient boosting models	Fully automated CVM classification pipeline capable of end-to-end prediction
RO4: To compare the performance of the automated CVM classification system with conventional visual CVM assessment by calibrated experts	Phase 3 Implementation and Validation	Model evaluation using accuracy, kappa statistics, and confusion matrices; comparison with expert consensus performance	Objective benchmarking of AI system against human expert assessment

Figure 3.1 Three-Phase Research Framework Linking Research Questions, Objectives, Activities, and Outcomes.

3.2 Sample Size and Sampling Technique

3.2.1 Pilot Study 1: Feasibility and Effect-Size Estimation

To enhance the quality and efficiency of this large-scale study, and in recognition of the limited investigations within the Malaysian population correlating skeletal age with CVM methods, an initial pilot study was conducted on 20 archived lateral cephalograms. The objectives were to test the feasibility of applying Baccetti's CVM staging method and to obtain a preliminary effect size for sample size determination. A correlation analysis between CVM stage and skeletal age produced a weak negative correlation ($r = –0.159$). Although expected due to the small sample size, this estimate provided a conservative basis for sample size planning, as weaker effect sizes demand larger samples for adequate statistical power. To avoid potential bias or duplication, the 20 pilot radiographs were excluded from the main study dataset (Magalhães et al., 2022).

3.2.2 Sample Size Calculation

Sample size was determined using the correlation coefficient from the pilot study and the standard Fisher z-transformation approach as applied in this recent study, as shown in Equation 3.1 (Mohammad-Rahimi et al., 2022):

$$N = \left(\frac{Z_\alpha + Z_\beta}{C} \right)^2 + 3 \tag{3.1}$$

Where:

- Correlation coefficient (r) = –0.159 (absolute value used for calculation)
- Significance level (α) = 0.05
- Power of study ($1–\beta$) = 0.80
- The standard normal deviate for $\alpha = Z\alpha = 1.9600$
- The standard normal deviate for $\beta = Z\beta = 0.8416$
- $C = 0.5 \times \ln[(1 + r) / (1 – r)] = 0.1604$

Based on these parameters, the required sample size was estimated at 900 radiographs (150 cases in each of six CVM stages).

While the statistical calculation indicated 900 samples, the development of a deep learning–based automated system necessitates larger datasets to achieve stable training and generalisable performance. Previous studies have demonstrated that neural network accuracy improves substantially with increasing dataset size (Tang et al., 2018). Therefore, a target sample size up to 1,200 radiographs (approximately 200 per CVM stage) was planned to support model development. Figure 3.2 illustrates the impact of data available on performance of traditional machine learning algorithms (that rely on hand-crafted features) and neural networks with few (shallow), moderate (medium), or large (deep) numbers of layers (Tang et al., 2018). Archived cephalograms were retrieved from institutional databases between 1 October 2023 and 1 April 2024**.**

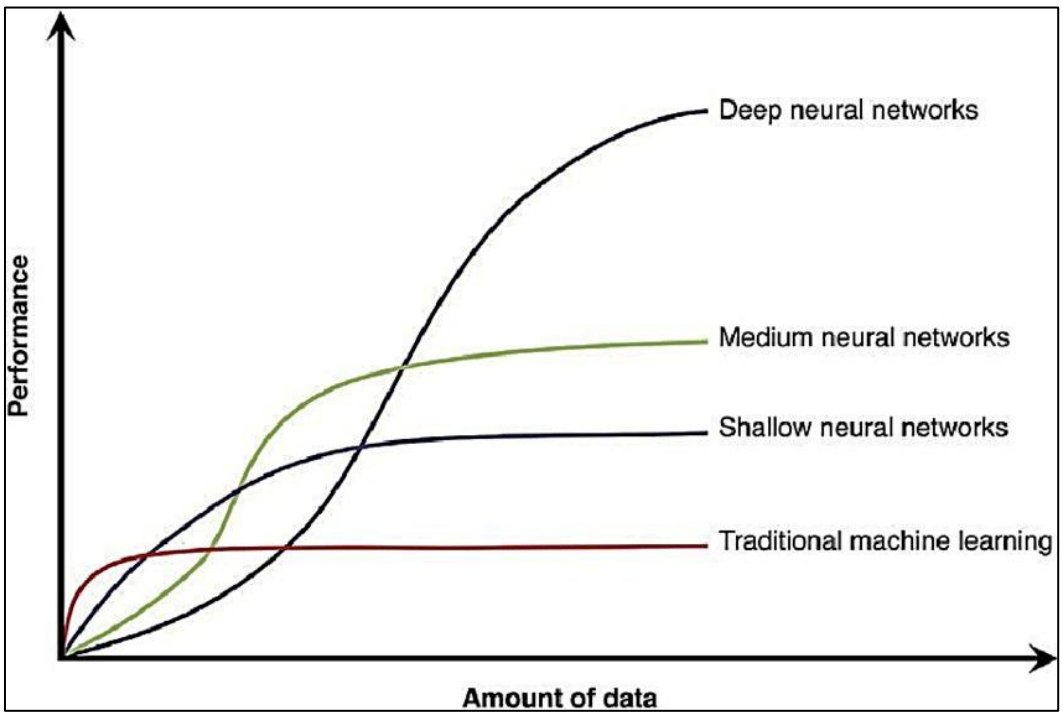


Figure 3.2 Effect of Data Size on Traditional Machine Learning and Neural Networks.

3.2.3 Sample Selection Criteria

Inclusion Criteria

Lateral cephalograms were eligible for inclusion if they met the following criteria:

1. Digitised LC images of Malaysian patients aged between 6 and 17 years.
2. Patients with no history of systemic conditions or medications known to significantly affect skeletal growth and bone metabolism.

3. Radiographs of sufficient quality, with the second (C2), third (C3), and fourth (C4) cervical vertebrae clearly visible.

Exclusion Criteria

Radiographs were excluded if they involved:

1. Patients with craniofacial syndromes, metabolic bone disorders, nutritional disorders affecting skeletal development, or history of medications known to alter bone metabolism (e.g., bisphosphonates).
2. Nonstandard head positioning or significant patient movement during image acquisition.
3. Presence of jewellery, retainers, or orthodontic appliances that interfered with image clarity.
4. Images demonstrating severe blurring, radiographic noise, or artefacts that obscured visualisation of the cervical vertebrae

3.3 Phase 1: Preliminary Study

3.3.1 Reliability Framework and Ground Truth Development

Phase 1 addressed Research Questions 1 and 2 by evaluating inter- and intra-observer reliability of CVM staging and establishing a calibrated ground truth dataset. This phase focused on assessing the consistency of CVM classification among orthodontic experts prior to automated model development, recognising that reliable expert labelling is essential for valid artificial intelligence training. Structured expert calibration, repeated assessments, and consensus adjudication were implemented to minimise subjectivity and standardise morphological interpretation. The outputs of Phase 1 comprised quantified reliability metrics and a consensus-labelled CVM dataset, which formed the reference standard for subsequent model development and validation phases.

3.3.2 CVM Classification

Skeletal maturation in this study was assessed using the CVM method proposed by Baccetti *et al.* (Baccetti et al., 2005). This method classifies skeletal development into six progressive stages based on characteristic morphological changes observed in the second (C2), third (C3), and fourth (C4) cervical vertebrae on lateral cephalometric radiographs. Figure 3.3 illustrates the six stages of the modified CVM method, demonstrating progressive morphological changes of C2–C4 from initiation (Stage I) through acceleration, transition, deceleration, maturation, and completion (Stage VI).

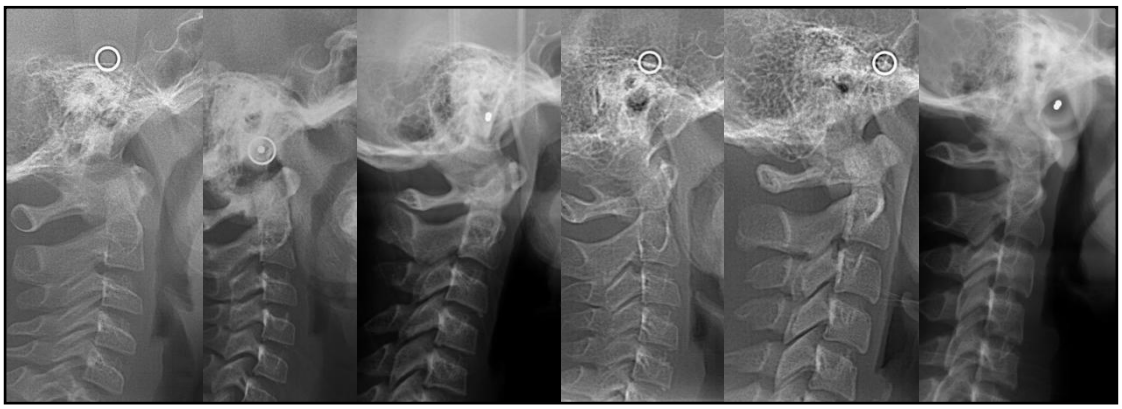


Figure 3.3 Six Stages of the Modified CVM Method

Morphological Criteria for CVM Stages

For this study, ground truth annotations were based on a modified version of Baccetti's CVM method, standardised into six operational stages (Figure 3.3):

- **Stage I**: Flat inferior borders of C2–C4; trapezoidal C3 and C4.
- **Stage II**: Inferior concavity at C2; C3 and C4 trapezoidal.
- **Stage III**: Concavities at C2 and C3; C3 becomes horizontally rectangular.
- **Stage IV**: Concavities at C2–C4; C3 and C4 horizontally rectangular.
- **Stage V**: Pronounced concavities; C3 and C4 square-shaped.
- **Stage VI**: Deep concavities; vertically rectangular C3 and C4 with elongated posterior borders.

This framework guided expert examiners in consistent and reproducible CVM staging for reliability analysis and model training.

3.4 Calibration Protocol and Reliability Assessment

Inter- and intra-observer reliability analyses were conducted to evaluate the consistency and reproducibility of CVM staging during expert calibration and dataset annotation. Reliability was assessed at multiple levels using Fleiss' weighted kappa for multi-expert agreement, Cohen's weighted kappa for agreement between two raters, and the ICC for intra-observer reproducibility (Fleiss, 1971; Fleiss et al., 1969). Details of each reliability assessment are described in the following subsections.

3.4.1 Establishing the Preliminary Reference Standard

Prior to multi-expert calibration, a preliminary reference set was established to provide an initial anchor for CVM staging and to reduce uncontrolled variability during subsequent calibration. The reference set consisted of 60 stratified lateral cephalometric radiographs, with equal representation of all six CVM stages.

Two experienced orthodontists independently staged the radiographs using predefined morphological criteria. Discrepant classifications were resolved through structured discussion to form a consensus-based preliminary reference standard, which was not disclosed to the expert panel during independent assessments.

3.4.2 Multi-Expert Assessment and Calibration

Following the establishment of the preliminary reference standard, a panel of thirteen orthodontic experts from diverse academic, governmental, and private practice backgrounds independently evaluated the same stratified set of 60 radiographs under blinded conditions. Orthodontic experts included registered orthodontists and academic specialists with a minimum of five years of clinical and/or academic experience in orthodontics following specialist qualification. Fleiss' weighted kappa was calculated following the initial assessment to establish baseline agreement. Structured consensus sessions were then conducted to refine ambiguous morphological criteria, particularly for transitional stages. After a one-month washout period, experts repeated the assessment in a second blinded round using the refined classification framework, with randomisation (Weir, 2005). Post-calibration agreement was quantified using Fleiss'

weighted kappa, and the refined criteria were adopted for subsequent analyses (Nestman et al., 2011; Perinetti et al., 2012).

Transitional intervals between CVM Stages II–III and III–IV exhibited lower sample prevalence and greater morphological ambiguity, contributing to inherent class imbalance and annotation uncertainty that reflect biological maturation rather than dataset artefact. During calibration, the expert panel identified recurring ambiguity in transitional CVM stages. To standardise adjudication, a structured operational decision protocol was introduced and applied in subsequent rounds. For cases appearing intermediate between CVM Stages II and III, concavity at C3 was compared relative to C2: less pronounced concavity at C3 indicated earlier maturation, while equal or greater concavity indicated later maturation. For cases appearing intermediate between CVM Stages III and IV, examiners assessed the inferior border of C5: a flat border indicated earlier maturation, whereas the presence of concavity indicated later maturation. Cases that remained unresolved after application of the protocol were escalated for structured group discussion and consensus classification. The protocol was used in subsequent annotation rounds to improve consistency during dataset labelling.

3.4.3 Inter-Observer Reliability Between Operational Examiners

Following calibration, the two orthodontic experts demonstrating the highest agreement and consistency during the multi-expert assessment phase were designated as operational examiners for full dataset annotation to optimise labelling consistency and reduce inter-annotator variability throughout large-scale dataset development. Inter-observer reliability between the two examiners was quantified using Cohen's weighted kappa, appropriate for assessing agreement between two raters assigning ordinal categorical data.

3.4.4 Intra-Observer Reliability (Intraclass Correlation Coefficient)

Each operational examiner re-assessed a randomly selected 15% subset of radiographs after a one-month washout period to minimise recall bias. Reproducibility was quantified using the ICC, employing a two-way random-effects model with absolute agreement. This analysis assessed the consistency of CVM staging by the same examiner across repeated evaluations.

3.5 Phase 2: Model Development

Phase 2 addressed Research Questions 3 and 4 through the development and evaluation of automated deep learning-based models for CVM classification. The CVM staging task in this study was formulated as an ordinal classification problem, reflecting the inherent ordered progression of skeletal maturation rather than a nominal multi-class structure. Building upon the calibrated ground truth dataset established in Phase 1, this phase aimed to investigate model feasibility, optimise classification performance, and compare automated predictions with expert-based CVM assessment.

Model development was conducted in two sequential stages. A pilot study employing ROS and a custom DCNN was first undertaken to establish baseline performance and identify modelling limitations. Findings from the pilot study informed the design of the final DCNN system, which adopted a two-stage architecture integrating YOLO-based vertebral localisation and Light Gradient Boosting Machine (LightGBM) classification.

3.5.1 Pilot Study 2: ROS-Based Custom DCNN for Feasibility Assessment

A pilot study was conducted to evaluate the feasibility of automated CVM classification using a custom DCNN trained on non-annotated lateral cephalometric radiographs. A custom DCNN was developed from scratch, without the use of pretrained weights or transfer learning, to establish an unbiased baseline pilot model for CVM stage classification. The purpose of this pilot phase was methodological rather than confirmatory, aiming to establish a baseline automated classification workflow, examine the effects of class imbalance, and identify architectural and training constraints prior to development of the final framework.

In the first stage, the custom DCNN was trained on the original imbalanced dataset without the application of ROS to assess baseline learning behaviour under natural CVM stage distributions. In the second stage, ROS was applied to the training dataset to balance class representation and to evaluate its effect on model training stability. In parallel, the performance of the custom DCNN was benchmarked against established pretrained convolutional neural networks, including InceptionV3, ResNet50, and MobileNetV2, to contextualise feasibility relative to commonly used transfer learning architectures.

The dataset was partitioned into training (80%) and testing (20%) subsets using stratified sampling to preserve the original CVM stage distribution. When applied, ROS was restricted exclusively to the training subset, while the validation and test sets remained unchanged to allow unbiased evaluation. Image preprocessing steps included pixel intensity normalisation, label encoding, and dataset integrity verification, implemented using NumPy and Pandas.

All preprocessing and model training were performed on a Dell Precision 5690 workstation equipped with an Intel Core Ultra 9 185H processor (2.50 GHz), 32 GB RAM, and an NVIDIA GPU. A custom DCNN architecture was developed and optimised during this pilot phase. Hyperparameter optimisation was conducted using random search, with parameters randomly sampled from predefined ranges including learning rate (0.0001–0.1), batch size (16–128), number of units in fully connected layers (64–512), dropout rate (0.2–0.5), number of convolutional filters (32–256), and kernel sizes (3×3 and 5×5). Each configuration was trained and evaluated based on validation performance. The predefined hyperparameter ranges were selected based on commonly reported configurations in deep learning image-classification literature and preliminary pilot experimentation to balance model complexity, convergence stability, and computational feasibility (Chollet, 2015).

Based on the selected configuration, the pilot model architecture comprised four convolutional layers with optimised filter sizes, each followed by batch normalisation and max-pooling layers. Feature aggregation was performed using a global average pooling layer, followed by a fully connected dense layer with ReLU activation and a dropout layer (rate = 0.2). The output layer consisted of six neurons corresponding to the six CVM stages, with softmax activation.

The model was implemented using TensorFlow/Keras and trained using the Adam optimiser with categorical cross-entropy loss. Training was conducted for up to 100 epochs using shuffled mini-batches, with early stopping applied based on validation performance. Model weights corresponding to the lowest validation loss were retained. To ensure independence between experiments, model weights were reinitialised and the computational graph was cleared prior to each training run using the Keras backend clear-session function (Mohammad et al., 2022; Norman, Mohd Rosli, et al., 2025).

3.5.2 Final Automated CVM Classification System

Building upon the reliability framework and pilot study, the final two-stage automated CVM classification system, termed CVMaXX, was developed based on the calibrated ground truth established in Phase 1. The system was implemented as a fully automated deep learning pipeline designed to replicate and enhance expert clinical decision-making.

3.5.2.1 Data Preprocessing, Image Annotation, and Dataset Preparation

All lateral cephalometric radiographs underwent a standardised preprocessing and annotation pipeline prior to model training. Images were imported, verified, and processed as grayscale arrays using OpenCV.

Manual image annotation was performed to generate ground truth labels for model training. LabelImg (version 1.8.6), a graphical image annotation tool was employed for lightweight bounding box annotation and were manually drawn around the second, third, and fourth cervical vertebrae (C2, C3, and C4) in each radiograph, following standardized anatomical boundaries. All annotations were reviewed to ensure consistency and accuracy, and annotation files were exported in YOLO text format (.txt), containing normalised class indices and bounding box coordinates compatible with the detection framework.

Fifteen percent of all images were randomly selected for re-annotation after a month interval to assess intra-observer reliability, following established protocols for medical image annotation studies (Kottner et al., 2011). Regular calibration sessions were conducted between annotators at two-week intervals during the annotation period to address emerging discrepancies and maintain consistency in interpretation criteria (Hallgren, 2012).

Bounding box annotation guidelines were defined based on anatomical landmarks and standardised radiographic interpretation principles. For the C2 vertebra, annotators were instructed to include the entire vertebral body from the superior to inferior cortical borders to ensure capture of all morphologically relevant structures (Hassel & Farman, 1995) (Figure 3.4a,b,c). For C3, the complete vertebral body was annotated with clear margin definition while avoiding overlap with adjacent anatomical structures (San Román et al., 2002) (Figure 3.4d). Similarly, C4 annotations

encompassed the full vertebral body, with particular attention to preventing overlap that could compromise subsequent automated analysis (Caldas et al., 2010) (Figure 3.4d). Disagreements were resolved through consensus discussion and iterative refinement of annotation guidelines (Gwet, 2008).

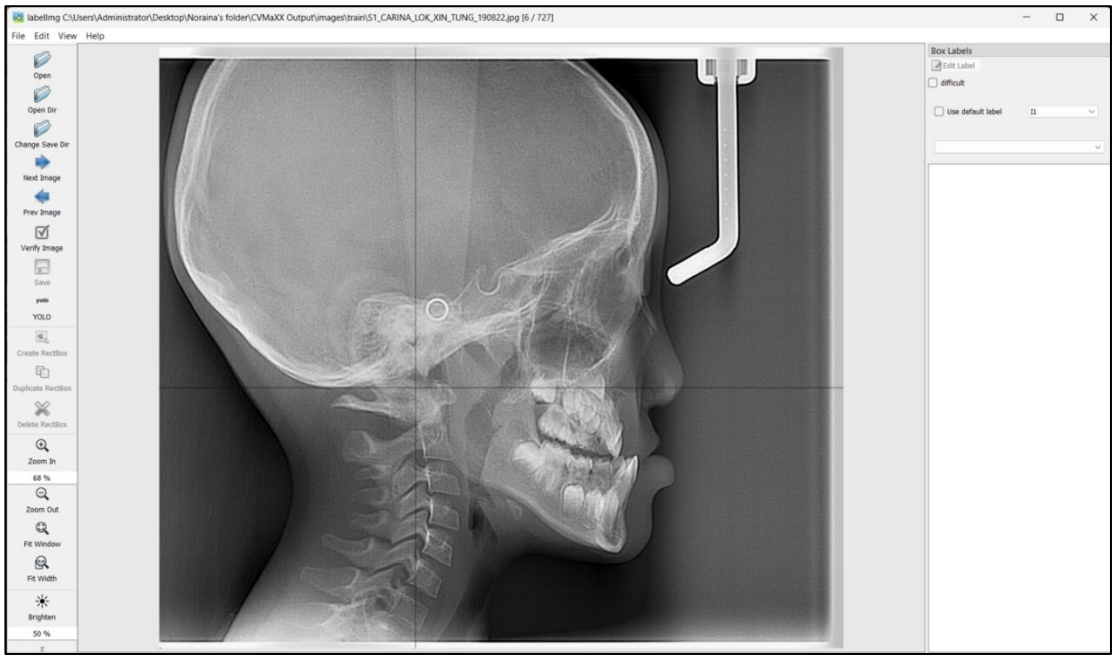


Figure 3.4a Manual Annotation of Cervical Vertebrae for Object Detection Model Training.

Representative screenshots illustrating the annotation workflow using LabelImg. Original lateral cephalometric radiograph prior to annotation.

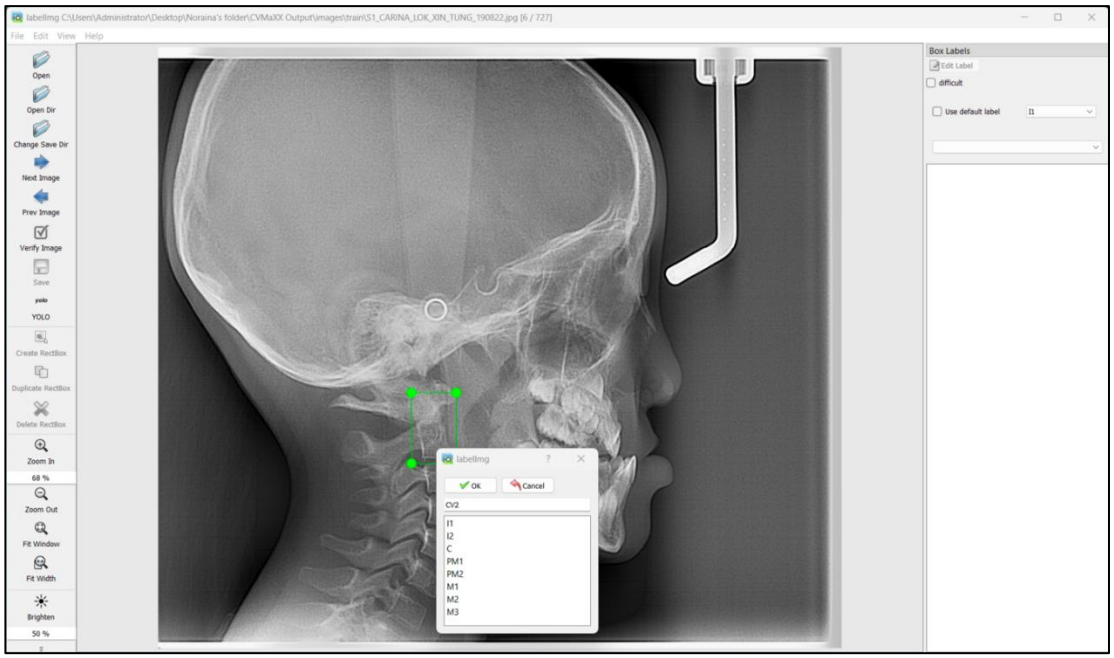


Figure 3.4b Assignment of Vertebral Labels During Annotation to Define Class-Specific Regions of Interest.

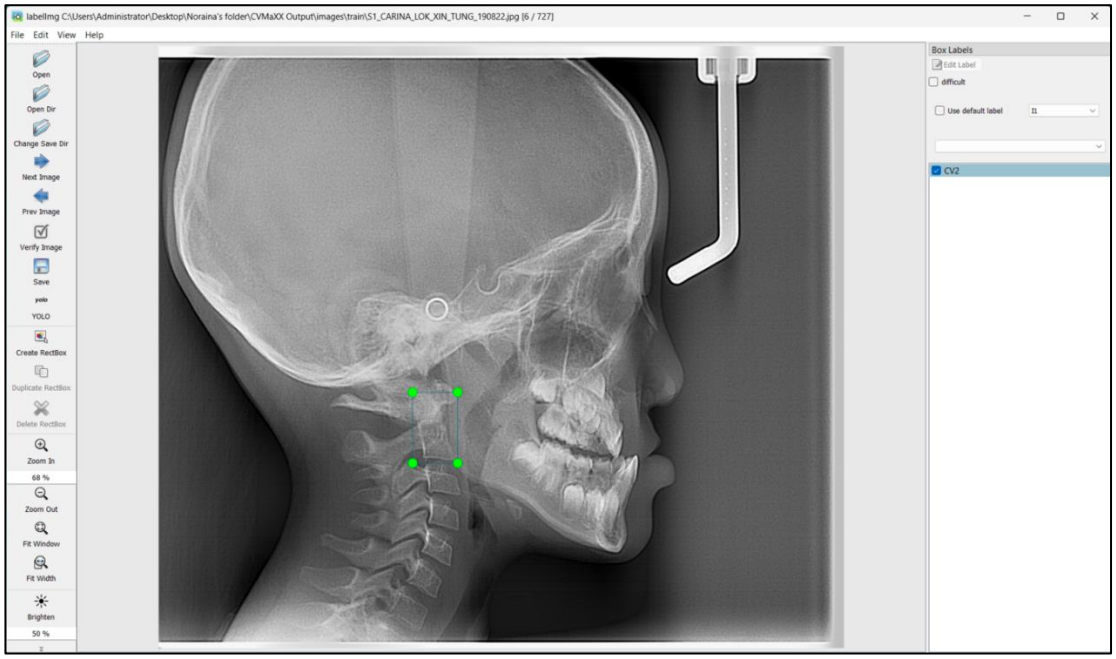


Figure 3.4c Bounding Box Annotation of the C2 Vertebral Body (CV2).

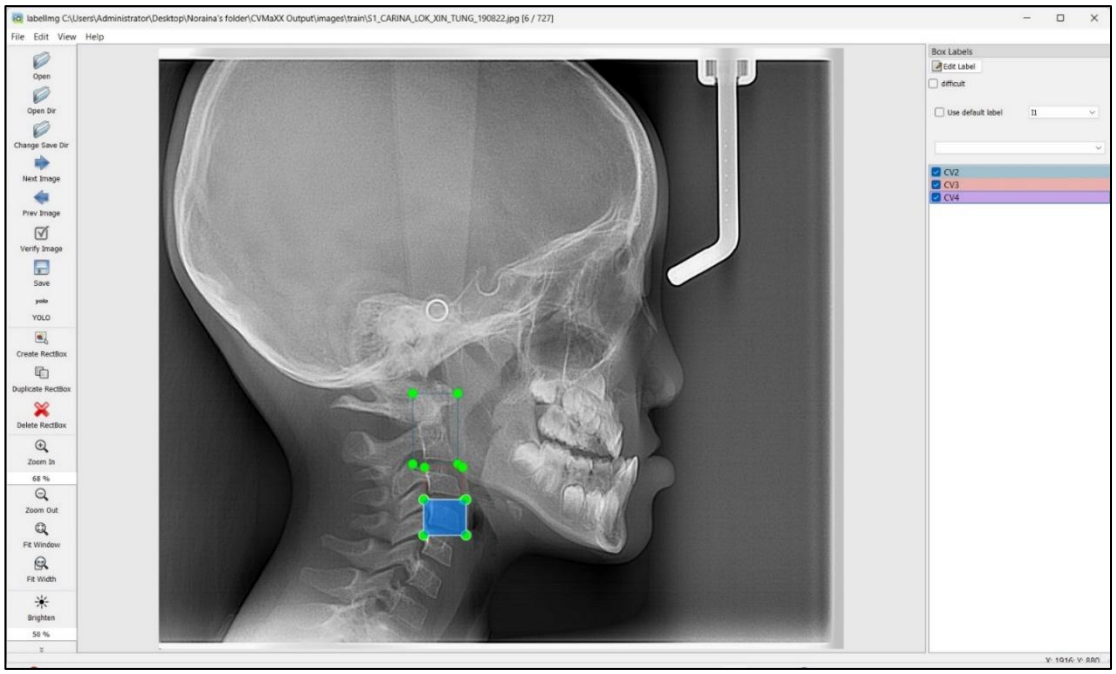


Figure 3.4d Final Annotated Image Showing Bounding Boxes for C2–C4 Vertebrae, used as Ground Truth for Training the YOLOv8 Object Detection Model in the CVMaXX System.

3.5.2.2 Stage 1: Object Detection

A localisation-first architecture was adopted to isolate C2–C4 vertebrae prior to staging, based on the clinical observation that diagnostic cues are spatially concentrated within these structures and may be obscured by surrounding anatomical variation. This approach also enhances interpretability by aligning model attention with clinician-derived regions of interest. Building on the standardised image annotation and dataset preparation procedures described in Section 3.6.1, bounding box annotations generated using LabelImg were used to train a YOLOv8 object detection model for automated localisation of the cervical vertebrae C2 to C4 (Jocher, 2023; Tzutalin, 2015). The YOLOv8 model was initialised with pretrained weights and subsequently fine-tuned on expert annotated lateral cephalometric radiographs, enabling accurate localisation of anatomically relevant vertebral regions for downstream analysis.

All annotations were exported in YOLO format for object detection tasks, while corresponding CSV files were retained to ensure label compatibility with the subsequent classification stage within the two-stage learning pipeline. The detection model employed a CSPDarknet53 backbone integrated with a Path Aggregation Network to enhance multiscale feature representation. Training was conducted for 300 epochs using the AdamW optimiser, with data augmentation strategies including rotation, scaling,

and photometric adjustments applied to improve robustness to variation in image quality and anatomical presentation.

3.5.2.3 Stage 2: Morphological Feature Extraction and Classifier Selection

Following YOLO-based localisation of the cervical vertebrae, morphological and textural features were extracted from the detected vertebral regions for CVM stage classification. Feature extraction was performed on the cropped regions corresponding to C2–C4 vertebrae to ensure that classification was based exclusively on anatomically relevant structures. The extracted feature set comprised geometric descriptors, boundary-based measurements, and texture-related features that reflect expert-defined morphological characteristics used in conventional CVM assessment.

The extracted features were subsequently used as inputs for supervised machine learning classification. Three ensemble-based classifiers were evaluated: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). These classifiers were selected due to their robustness in handling structured tabular data, ability to model non-linear feature interactions, and suitability for multi-class classification tasks involving heterogeneous feature sets.

For each classifier, hyperparameter optimisation was performed using randomised search combined with five-fold cross-validation on the training dataset. For the Random Forest model, tuned hyperparameters included the number of trees, maximum tree depth, and feature sampling strategy. For XGBoost, optimisation focused on the learning rate, maximum tree depth, subsampling ratios, and regularisation parameters, including L1 and L2 penalties. LightGBM was configured using a leaf-wise tree growth strategy with depth limitation, gradient-based one-side sampling, and feature bundling to improve computational efficiency and performance when handling high-dimensional morphological and textural features.

Classifier performance was evaluated using an independent test dataset through Receiver Operating Characteristic (ROC) curve analysis. ROC curves were generated for each classifier, and the Area Under the Curve (AUC) was calculated to quantify discriminative ability across CVM stages. Statistical comparisons between classifiers were conducted using the DeLong test for correlated ROC curves to assess differences in classification performance under identical testing conditions. Based on comparative

evaluation across AUC, sensitivity, and specificity measures, LightGBM was selected as the final classifier for integration into the CVMaXX automated classification system.

3.6 Phase 3: Performance Evaluation And Statistical Analysis

The final CVMaXX model was evaluated using an independent unseen testing dataset and benchmarked against expert derived ground truth annotations. Agreement between automated CVM classification and expert assessment was evaluated using unweighted Cohen's kappa, treating the automated system as an independent rater. This approach was selected to assess overall case-level agreement rather than the magnitude of ordinal disagreement between adjacent CVM stages.

Model performance was further assessed using standard classification metrics, including accuracy, precision, recall (sensitivity), specificity, and F1 score. These metrics were derived from confusion matrices generated across the six CVM stages to quantify overall classification performance and stage specific misclassification patterns. All statistical analyses were performed using IBM SPSS Statistics version 23.0.

3.6.1 Definition of Inclusive and Strict Evaluation Criteria

In addition to standard predictive performance metrics, descriptive statistical analysis was performed to compare automated CVM staging with expert derived reference stages. For each case, the numerical difference between the predicted CVM stage and the reference stage was calculated and summarised as mean staging error to characterise the direction and magnitude of deviation.

To reflect both clinically acceptable variability and strict diagnostic precision, model performance was evaluated using two complementary evaluation frameworks: an inclusive evaluation and a strict evaluation criterion.

The inclusive evaluation was designed to capture clinically acceptable staging variability. Under this criterion, a prediction was considered correct if it either exactly matched the expert assigned CVM stage (true positive, TP) or fell within one adjacent stage above or below the reference stage (true negative, TN). This approach acknowledges the biological continuity of skeletal maturation and the well documented transitional ambiguity between adjacent CVM stages. Predictions differing by more than one stage from the reference classification were considered incorrect.

Under the inclusive evaluation framework, accuracy was calculated as (Equation 3.2):

$$\text{Accuracy}_{\text{inclusive}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

where TP represents exact stage agreement, TN represents predictions within one adjacent CVM stage, FP represents incorrect predictions above one stage difference, and FN represents incorrect predictions below one stage difference.

The strict evaluation adopted a conservative criterion in which only exact stage agreement with the expert reference was considered correct. Under this framework, all non-exact predictions, including those within adjacent stages, were classified as incorrect. This approach provides a stringent assessment of model precision and reflects clinical scenarios where exact CVM staging is required.

Under the strict evaluation framework, accuracy was calculated as (Equation 3.3):

$$\text{Accuracy}_{\text{strict}} = \frac{TP}{TP + FP + FN} \tag{3.3}$$

where TP represents exact stage agreement only.

Agreement between automated CVM staging and expert assessment was quantified using unweighted Cohen's kappa (κ). This measure was selected to evaluate overall case-level agreement between the automated system and human examiners, with all disagreements treated equally regardless of the ordinal distance between CVM stages.

Given the ordinal nature of cervical vertebral maturation staging, weighted Cohen's kappa statistics were additionally computed using linear and quadratic weighting schemes to account for the magnitude of disagreement between predicted and reference stages. Quadratic weighted κ was emphasised, as it assigns greater penalties to larger stage discrepancies while allowing tolerance for minor adjacent stage disagreement, thereby providing clinically meaningful supplementary interpretation.

CHAPTER 4
RESULTS

4.1 Introduction

This chapter presents the outcomes of the model development, training, and evaluation phases described in Chapter 3. The results are organised into detection accuracy, classification performance, confusion matrix analysis, interobserver comparison, and model refinement outcomes. Figure 4.1 illustrates the experimental workflow of the CVMaXX system, providing an end-to-end overview from data acquisition and expert ground truth calibration to model development, internal evaluation, and external validation. The workflow also highlights how pilot study findings informed the final two-stage architecture, which integrates vertebral localisation with morphological feature-based classification.

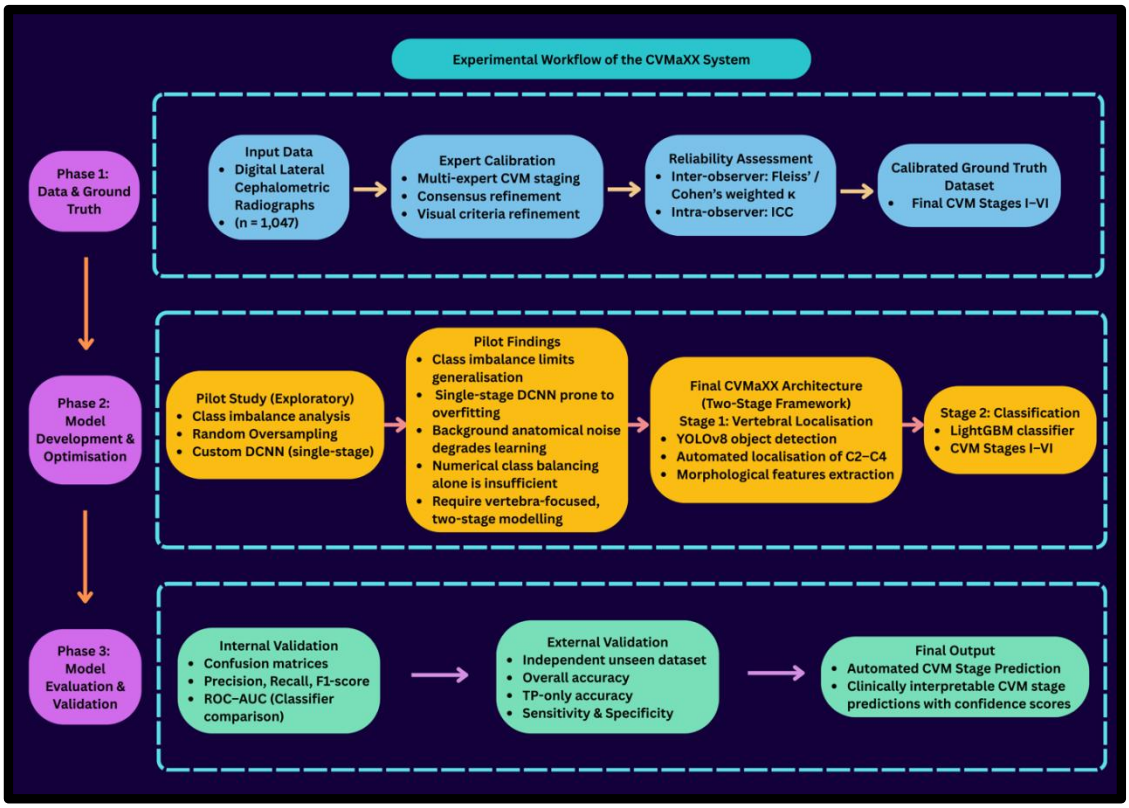


Figure 4.1 CVMaXX Model Development and Validation Workflow.

4.2 Phase 1: Sample Distribution

A total of 1609 lateral cephalometric radiographs were initially collected from six public universities and private clinics across Malaysia (Table 4.1). The institutional contributions varied: Universiti Teknologi MARA (UiTM) provided the largest share of radiographs (n = 556), followed by Universiti Malaya (UM, n = 386). Private clinics collectively contributed 581 radiographs, while smaller numbers came from Universiti Sains Malaysia (USM, n = 25), Universiti Kebangsaan Malaysia (UKM, n = 30), and Universiti Sains Islam Malaysia (USIM) (n = 31).

The distribution across CVM stages was uneven, with the largest number of radiographs in CVM Stage V (n = 425), CVM Stage IV (n = 373) and CVM Stage VI (339). Intermediate stages such as CVM Stage II (n = 152) and CVM Stage III (n = 225) were moderately represented, while CVM Stage I (n = 95) had the least samples.

To ensure balanced representation for automated model training and validation, the dataset was capped at a maximum of 200 radiographs per stage. After this adjustment, the final working dataset comprised 1,047 radiographs, with near-equal representation across the six CVM stages (200 per stage). This curated dataset was used for both reliability assessment and subsequent AI model development.

Table 4.1
Institutional Contribution

Institution	Stage I	Stage II	Stage III	Stage IV	Stage V	Stage VI	Total
UM	33	46	57	151	83	16	386
UiTM	44	66	96	49	117	184	556
USM	7	7	11	0	0	0	25
UKM	3	11	16	0	0	0	30
USIM	0	3	7	0	0	21	31
Private Clinics	8	19	38	173	225	118	581
Total	**95**	**152**	**225**	**373**	**425**	**339**	**1609**

4.3 Reliability Assessment

4.3.1 Establishing the Preliminary Gold Standard

Two senior orthodontists independently assessed 60 stratified lateral cephalometric radiographs, with 10 images representing each of the six CVM stages (Figure 4.2). Initial discrepancies between the examiners were resolved through discussion, producing a consensus-based preliminary gold standard. This dataset served as the reference for subsequent reliability testing by the multi-expert panel.

4.3.2 Multi-Expert Assessment and Calibration

Thirteen orthodontic experts participated in the reliability study, representing academic institutions (n = 10), a private academic institution (n = 1), government service (n = 1), and private practice (n = 1) (Table 4.2). Each expert independently staged the 60 radiographs under blinded, randomised conditions. Variability was most evident at transitional stages, particularly CVM Stage III. A structured consensus session was held to resolve these disagreements, resulting in refined operational definitions for ambiguous morphological features. After a one-month washout period, the experts repeated the assessments, enabling both inter- and intra-observer reliability to be estimated.

Table 4.2
Institutional Affiliation and Distribution of Orthodontic Expert Assessors Participating in the Multi-Expert Reliability Study.

Institutions	Assessors
Government Academic Institutions:	
UiTM	3
USM	1
UM	2
UKM	2
UIAM	1
USIM	1

Private Academic Institutions:	
MSU	1
Government Sector (Ministry of Health):	
Selangor	1
Private Practitioner:	
Selangor	1

4.3.3 Statistical Analysis

All statistical analyses were performed using SPSS version 28.0 (IBM Corp., Armonk, NY, USA). Inter-observer agreement among multiple raters was assessed using Fleiss' κ , while intra- and inter-observer agreement for operational examiners was assessed using Cohen's weighted κ and the ICC (Cohen, 1968; Fleiss, 1971; Fleiss et al., 1969). Interpretation followed Landis and Koch's criteria (Landis & Koch, 1977):

- $\kappa < 0.20$: Poor agreement
- $\kappa = 0.21$-0.40: Fair agreement
- $\kappa = 0.41$-0.60: Moderate agreement
- $\kappa = 0.61$-0.80: Substantial agreement
- $\kappa = 0.81$-1.00: Almost perfect agreement

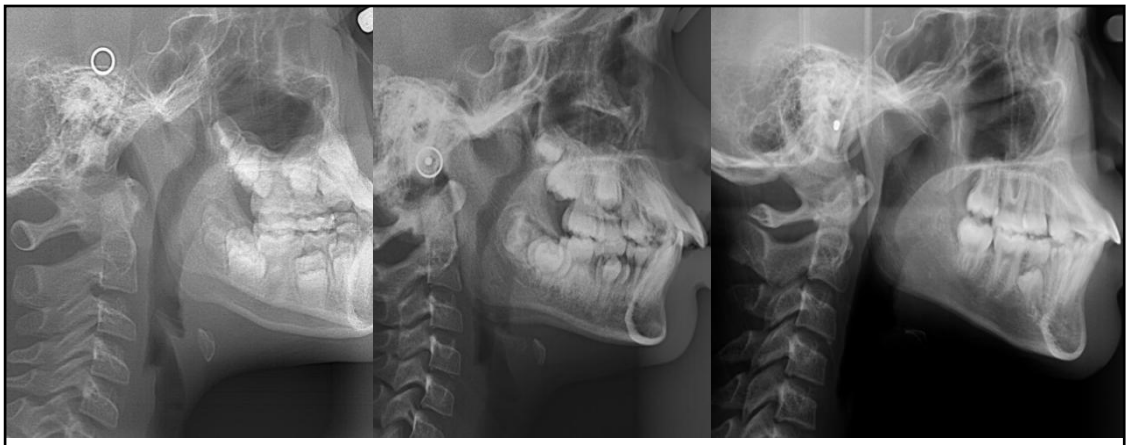
A significance threshold of $p < 0.05$ was applied. To improve precision, 95% confidence intervals were generated using bootstrap resampling methods.

In the first round of multi-expert assessments, Fleiss' weighted κ was 0.580 (95% CI: 0.465–0.657, $p < 0.001$), indicating moderate agreement across the 13-expert panel. Disagreements were concentrated in transitional stages (CVM Stage II, III and IV). Following structured calibration sessions and refinement of staging definitions, agreement improved in the second round, with Fleiss' weighted κ increasing to 0.611 (95% CI: 0.492–0.641), $p < 0.001$), representing substantial agreement.

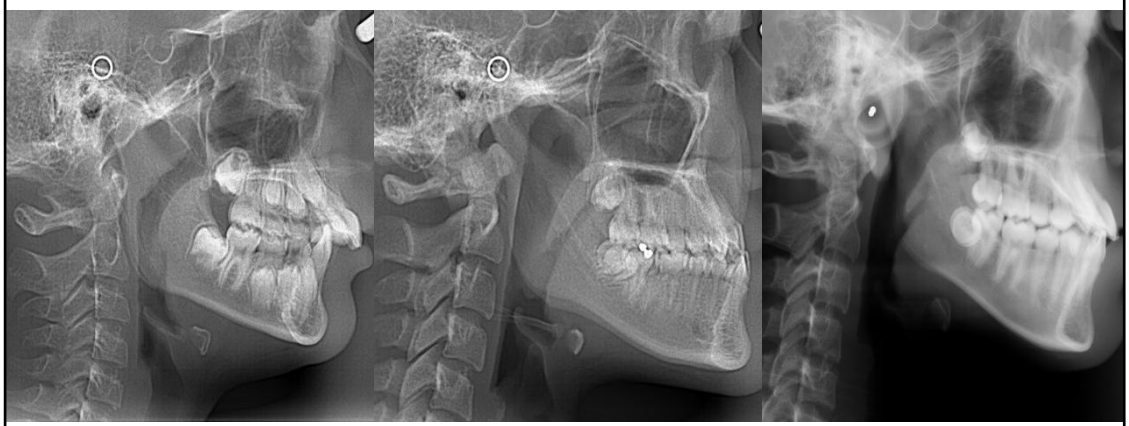
4.3.4 Operational Decision Protocol for Ambiguous Cases

During calibration, the expert panel identified recurrent ambiguity in transitional CVM stages. To standardise adjudication and facilitate consensus, a structured decision protocol was introduced and applied in subsequent rounds. The protocol consisted of stage-specific visual inspection criteria to resolve intermediate cases, as illustrated in Figures 4.3a–b. Persistently ambiguous cases were escalated for structured group discussion and consensus classification.

Following implementation, the protocol was applied to approximately 15% of cases, of which 89% achieved consensus without extended discussion. Application of the protocol reduced transitional ambiguity and was associated with an improvement in inter-examiner reliability, with Fleiss' weighted κ increasing from 0.580 to 0.611 across calibration rounds.



Stage I Stage II Stage III



Stage IV Stage V Stage VI

Figure 4.2 Six Stages of the Modified Baccetti CVM Method. Illustration of Progressive Morphological Changes in Cervical Vertebrae C2–C4 across CVM Stages I to VI, Corresponding to Initiation, Acceleration, Transition, Deceleration, Maturation, and Completion Phases of Pubertal Skeletal Growth.

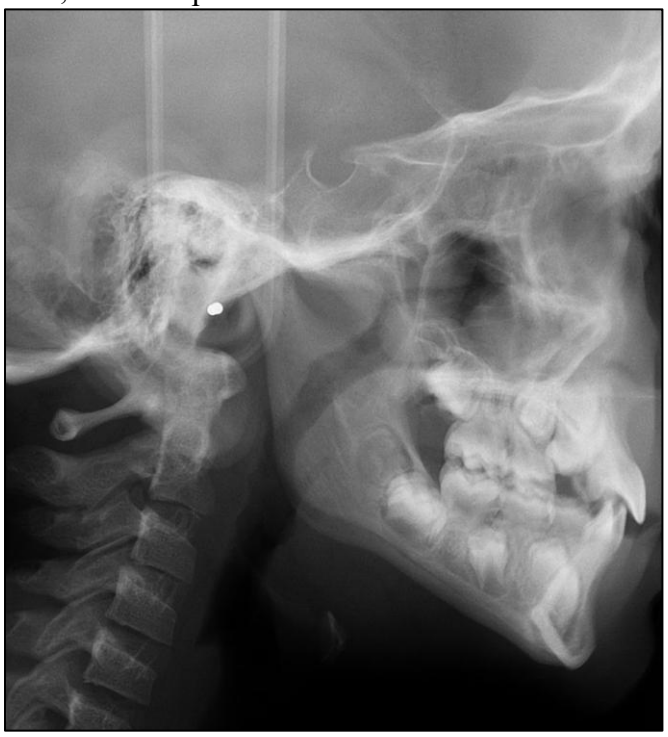


Figure 4.3a For Borderline Cases between CVM Stages II and III, Examiners Compared the C3 Curve to C2. A Shallow or Barely defined C3 Concavity Indicated Stage II, while a C3 Curve Approaching the C2 Concavity Indicated Stage III. C4 was Flat in Both Stages.

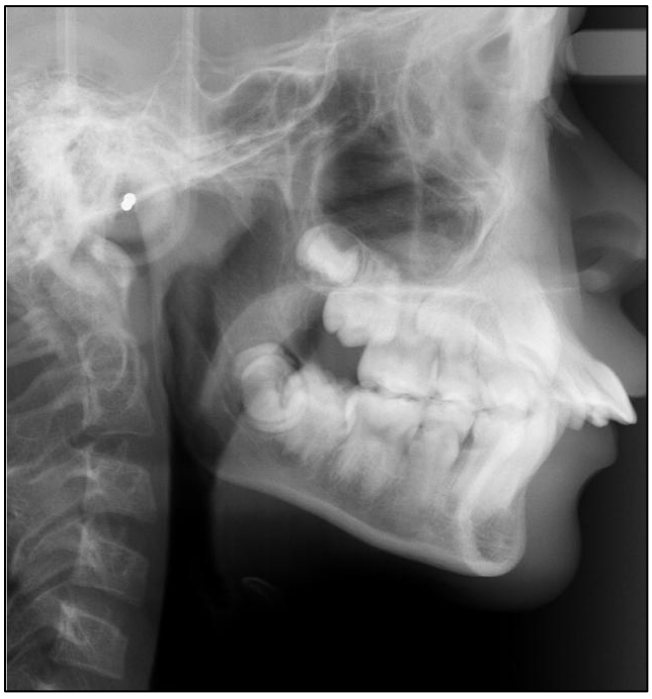


Figure 4.3b For Borderline Cases between CVM Stages III and IV, Examiners Inspected the Sequence of Inferior Border Concavity across the Lower Cervical

Vertebrae. In CVM Stage III, Concavity at C3 and C4 was Present while C5 Remained Flat, Whereas in CVM Stage IV, Concavity Extended to C5.

4.3.5 Intra-Observer Reliability

The two operational examiners demonstrated excellent temporal stability. The ICC was 0.973 (95% CI: 0.962–0.984), indicating outstanding absolute agreement. Cohen's κ was 0.870 (95% CI: 0.834–0.906), reflecting almost perfect categorical agreement. These results confirm high reproducibility and minimal diagnostic drift, providing a reliable benchmark for subsequent AI validation.

4.4 Phase 2: Pilot Model Performance without Random Oversampling

Phase 2 presents exploratory pilot experiments conducted to characterise baseline learning behaviour, assess the effects of class imbalance, and identify limitations of single-stage DCNN classification using non-localised lateral cephalometric radiographs. These analyses were undertaken for feasibility and diagnostic purposes only and were not used to train or optimise the final CVMaXX system.

4.4.1 Dataset Description

A total of 922 digital lateral cephalometric radiographs were used in this pilot phase, reflecting the original imbalanced distribution of CVM stages observed in routine clinical records. The custom DCNN was trained without ROS to establish a baseline reference under real-world class distribution conditions. Table 4.3 summarises the mean age and distribution of CVM stages based on expert visual assessment.

Table 4.3
Descriptive Statistics of the Patient's Age and Cervical Vertebrae Stage

Cervical vertebrae maturation stages	Mean age (Years) ±SD	N (%)
CVM Stage I	7.59 ± 1.64	32 (3.47%)

CVM Stage II	9.75 ± 1.53	47 (5.10%)
CVM Stage III	10.61 ± 1.24	97 (10.52%)
CVM Stage IV	12.16 ± 1.31	196 (21.26%)
CVM Stage V	13.62 ± 1.35	296 (32.10%)
CVM Stage VI	17.21 ± 3.04	254 (27.55%)
Total	11.82 ± 3.35	922 (100%)

4.4.2 Training and Validation Accuracy

The pilot DCNN achieved an apparent training accuracy of 100%, whereas validation accuracy remained substantially lower at 57%, indicating a pronounced discrepancy between training and validation performance.

4.4.3 Learning Curves and Error Patterns

Figure 4.4 illustrates the training and validation accuracy curves for the pilot model trained without ROS. Training accuracy increased rapidly and reached saturation, while validation accuracy plateaued at a markedly lower level. The divergence between the two curves was evident early in the training process.

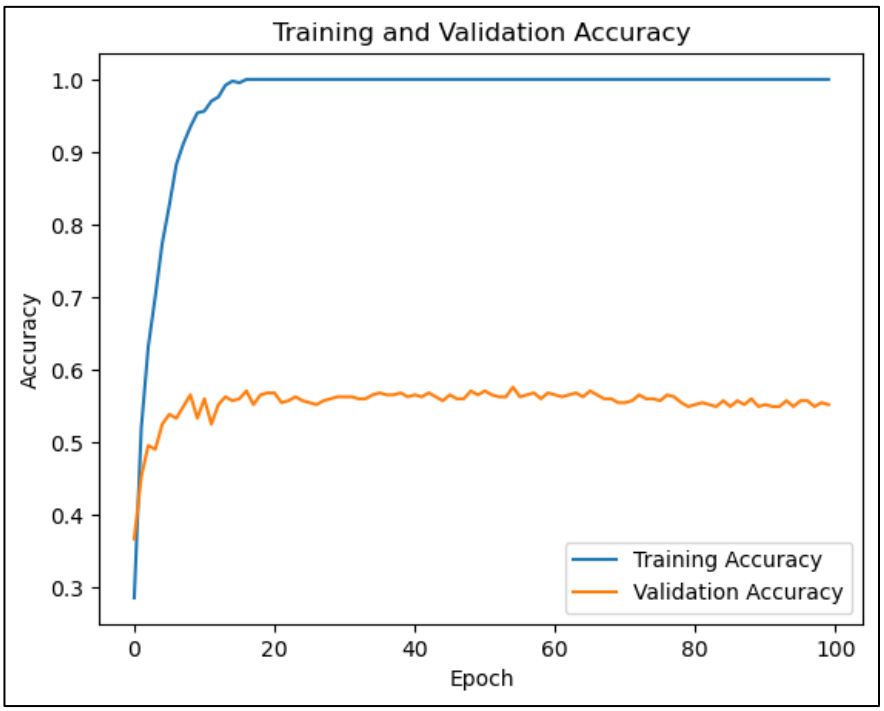


Figure 4.4 Learning Curves of the Pilot DCNN Trained without ROS Training and Validation Accuracy Trajectories Demonstrating Rapid Overfitting under Imbalanced Class Conditions, with Early Divergence between Training and Validation Performance.

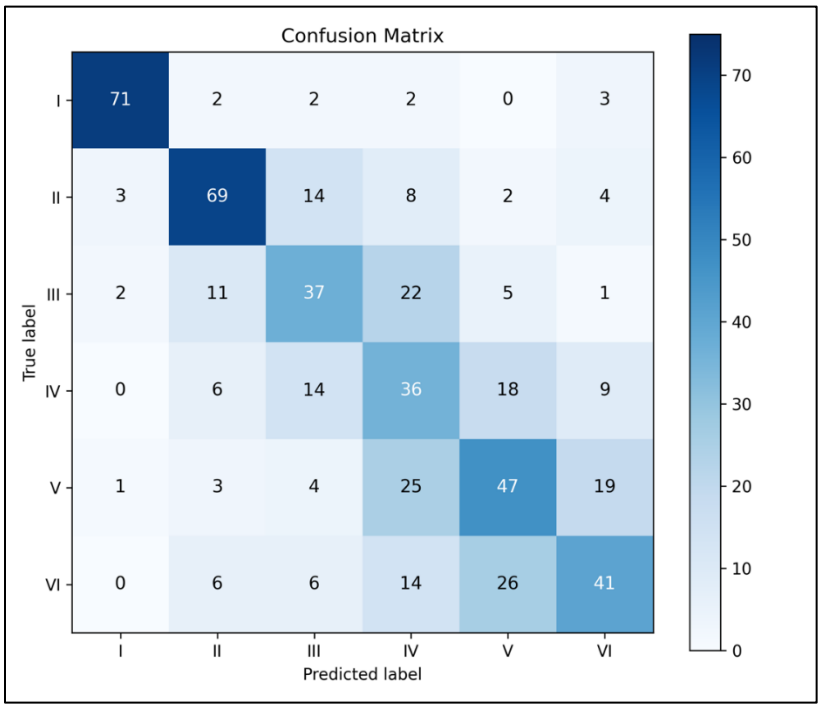


Figure 4.5 Confusion Matrix of the Pilot DCNN Trained without Random Oversampling.

Figure 4.5 presents the confusion matrix for the pilot model trained without ROS. Correct classifications were primarily observed along the diagonal; however, a high frequency of misclassifications was evident between adjacent CVM stages. Errors were most prominent among CVM Stages III through VI, with overlap observed between neighbouring stages.

4.5 Pilot Model Performance with Random Oversampling

Following baseline evaluation under imbalanced conditions, additional pilot experiments were conducted to examine the effect of random oversampling on classification stability and performance under balanced training conditions. These experiments were intended to explore the upper-bound capability of the pilot DCNN architecture when class imbalance was mitigated, rather than to establish a deployable model.

4.5.1 Oversampling Strategy and Balanced Dataset Size

ROS was applied only to the training subset, increasing representation of underrepresented CVM stages while preserving the original class distributions in the validation and test sets. Initially, class sizes ranged from 32 images in CVM Stage I to 296 images in CVM Stage V. After oversampling, each class in the training subset was expanded to 296 images, resulting in 1,420 images for training and 356 images for validation.

This oversampling procedure accounts for differences in reported image counts across sections of the pilot experiments. While the original dataset comprised 922 images, the oversampled training and validation subsets exceeded 1,700 images due to duplication during ROS.

Hyperparameter optimisation was subsequently performed to identify a stable configuration for the pilot architecture. It identified an optimal configuration consisting of 112 units in the fully connected dense layer and a dropout rate of 0.2. This configuration was selected for the final model and applied consistently across all subsequent training and evaluation experiments (Figure 4.6).

Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 126, 126, 96)	2688
max_pooling2d_2 (MaxPooling 2D)	(None, 63, 63, 96)	0
conv2d_3 (Conv2D)	(None, 61, 61, 112)	96880
max_pooling2d_3 (MaxPooling 2D)	(None, 30, 30, 112)	0
flatten_1 (Flatten)	(None, 100800)	0
dense_2 (Dense)	(None, 112)	11289712
dropout_1 (Dropout)	(None, 112)	0
dense_3 (Dense)	(None, 6)	678

Total params: 11,389,958
Trainable params: 11,389,958
Non-trainable params: 0

Figure 4.6 Detailed Layer Configuration of the Pilot Model.

4.5.2 Training Performance

During the training of the pilot model over 100 epochs, significant improvements in performance were observed. Initially, the model achieved a loss of 1.7053 and an accuracy of 33.24% on the training set, with a validation accuracy of 50.56%. By the fifth epoch, training accuracy had increased to 85.92%, with validation accuracy rising to 79.78%. As training progressed, accuracy continued to improve, reaching 97.39% by epoch 10, although validation accuracy experienced fluctuations, peaking at 87.36% in epoch 15. After the 50th epoch, training was stopped due to the early stopping function. Despite some variability in validation accuracy, ranging between 82.30% and 87.36%, the model achieved a best-performing pilot configuration of 88.2% by epoch 50. Figure 4.7 shows the learning curve for classifying CVM stages, illustrating the model's performance.

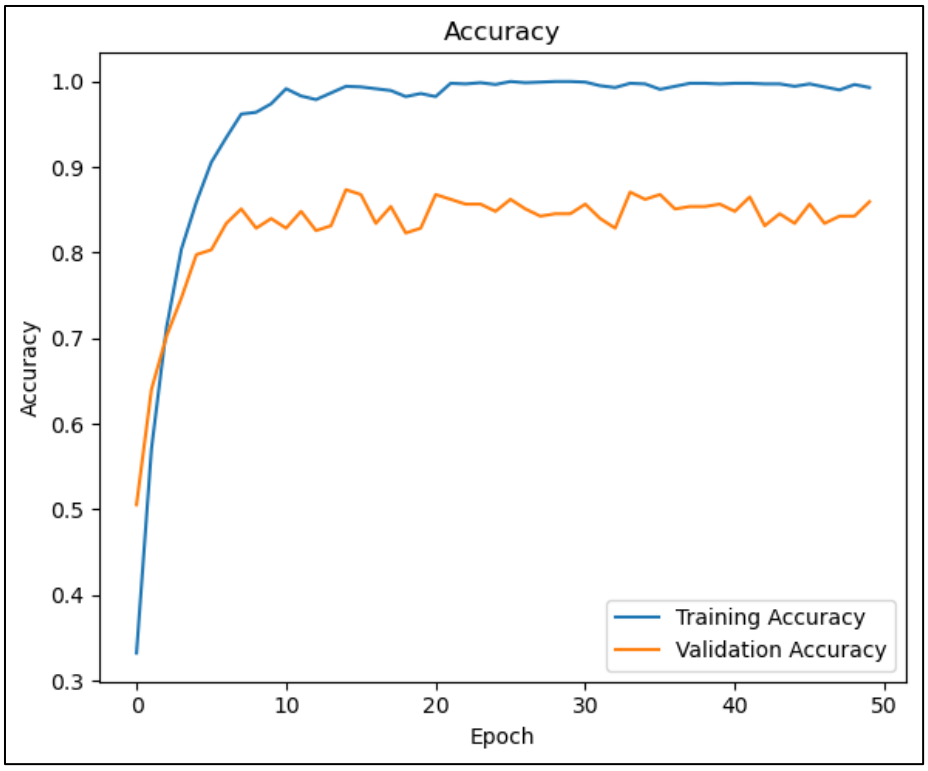


Figure 4.7 Training and Validation Accuracy during Pilot DCNN Training with ROS.

4.5.3 Classification Performance

Performance patterns across CVM stages were first examined using the confusion matrix (Figure 4.8), which illustrates class-wise prediction behaviour of the

pilot model following random oversampling. The model demonstrated strongest performance in the early stages, with high correct classification rates for CVM Stages I–III. Misclassifications in these stages were minimal and largely confined to adjacent stages. In contrast, reduced performance was observed in the later stages, particularly CVM Stage V, which exhibited a higher degree of confusion with neighbouring stages. CVM Stage IV also showed moderate misclassification, while CVM Stage VI demonstrated intermediate performance with both correct and incorrect predictions distributed across adjacent categories. Overall, the confusion matrix highlights a clear trend of increasing classification difficulty in the later and transitional CVM stages.

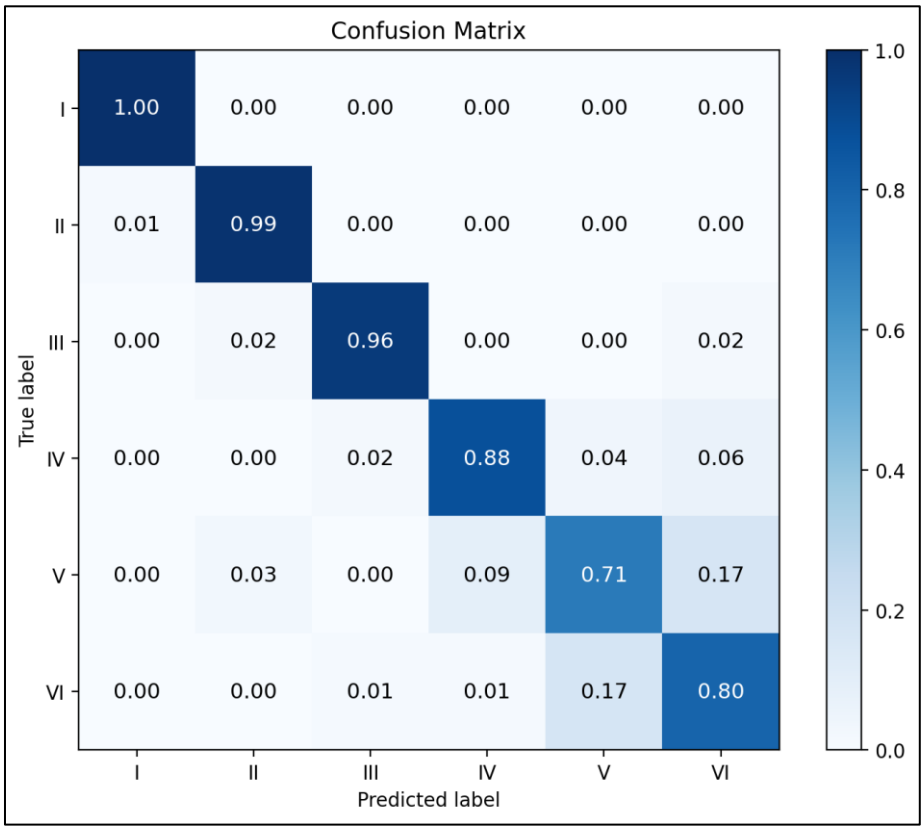


Figure 4.8 Confusion Matrix of the Pilot CVM Classification Model after ROS.

Table 4.4
Precision, Recall, and F1-Score Values for each CVM Stage (I–VI) Calculated from the Confusion Matrix.

Stage	Precision	Recall	F1- score
I	0.980	1.00	0.990
II	0.960	0.990	0.970
III	0.960	0.960	0.960
IV	0.857	0.875	0.870
V	0.770	0.712	0.740
VI	0.790	0.800	0.790
Accuracy	0.882	0.882	0.882

Quantitative performance metrics derived from the confusion matrix are summarised in Table 4.4. Precision and recall were highest for CVM Stages I–III, with precision values of 0.98, 0.96, and 0.96 and recall values of 1.00, 0.99, and 0.96, respectively. CVM Stage IV showed slightly lower performance, with precision of 0.857 and recall of 0.875. The weakest performance was observed for CVM Stage V, which recorded a precision of 0.770 and recall of 0.712, indicating greater difficulty in accurately distinguishing this stage. CVM Stage VI achieved moderate performance, with precision of 0.790 and recall of 0.800. The corresponding F1-scores followed similar trends across stages.

Overall, the pilot model achieved a classification accuracy of 88.2%. Macro- and weighted-average metrics further indicate balanced performance across CVM stages, with the macro average reflecting class-level consistency and the weighted average accounting for class imbalance. Together, the confusion matrix (Figure 4.8) and performance metrics (Table 4.4) demonstrate strong performance in early CVM stages and highlight persistent challenges in later and transitional stages.

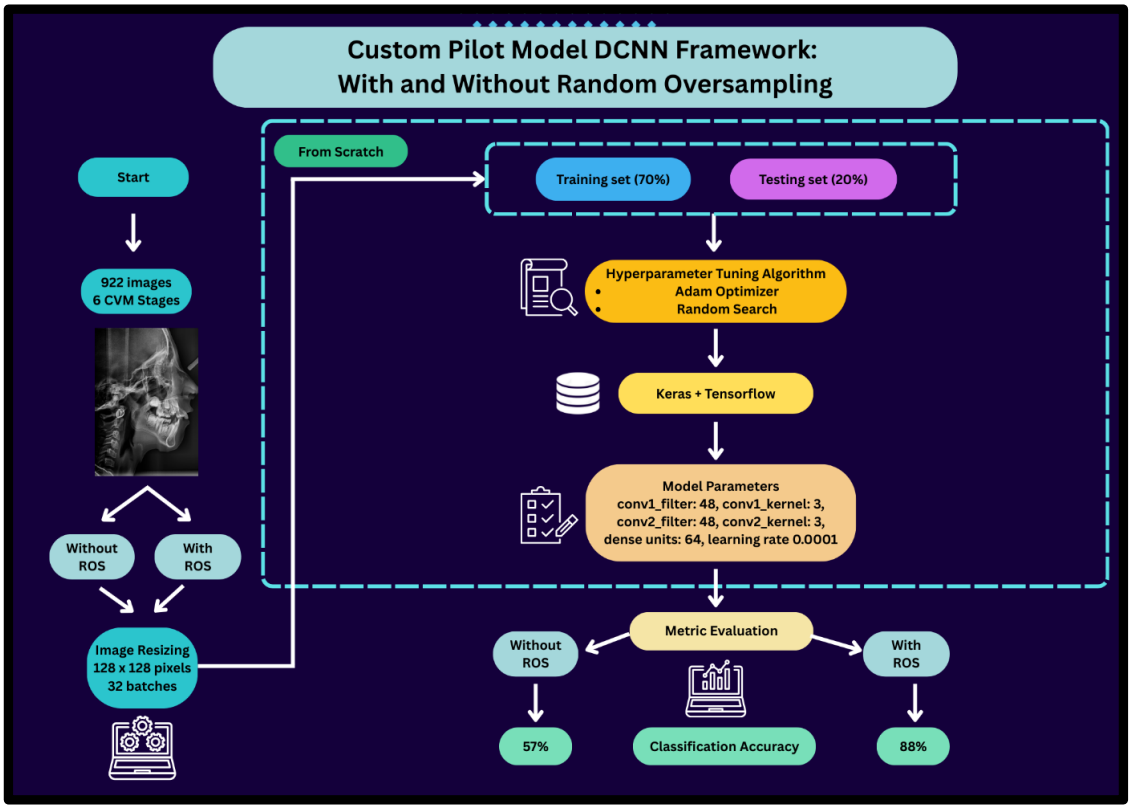


Figure 4.9 Workflow of the Pilot DCNN Model Development and Evaluation. Illustration of Dataset Preparation with and without Random Oversampling, Image Preprocessing, Model Training, Hyperparameter Tuning, and Comparative Performance Assessment.

Figure 4.9 demonstrates the overall workflow of the custom pilot DCNN model development and evaluation for CVM stage classification. A total of 922 lateral cephalometric images representing six CVM stages were included. To assess the effect of class imbalance handling, the dataset was processed under two conditions: without ROS and with ROS. All images were resized to 128 × 128 pixels and trained using a batch size of 32. The dataset was split into a training set (80%) and a testing set (20%). Hyperparameter tuning was performed using a random search strategy with the Adam optimizer implemented in Keras and TensorFlow. The final model architecture consisted of two convolutional layers followed by a dense layer, with key parameters including 48 filters per convolutional layer, a kernel size of 3, 64 dense units, and a learning rate of 0.0001. Model performance was evaluated using classification accuracy, demonstrating a marked improvement when ROS was applied (88%) compared with the non-oversampled model (57%).

4.5.4 Pretrained Model Comparison

Figure 4.10 compares the pilot model with selected pretrained architectures (InceptionV3, ResNet50, and MobileNetV2) under identical experimental settings. The proposed pilot model achieved the highest training accuracy and maintained a stable validation accuracy above 85%. In contrast, InceptionV3 demonstrated lower training accuracy with fluctuating validation performance, while ResNet50 showed consistently poor performance. MobileNetV2 performed better than InceptionV3 and ResNet50, but validation accuracy remained more variable than the pilot model.

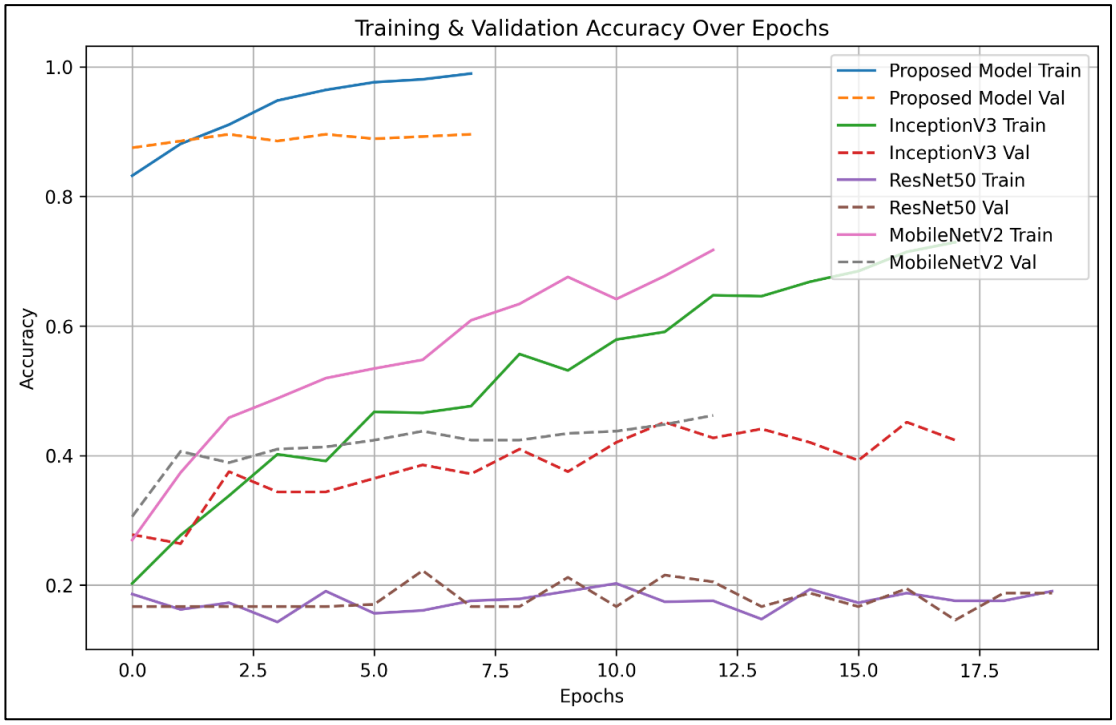


Figure 4.10 Performance Comparison between the Custom Pilot DCNN and Selected Pretrained Architectures under Identical Experimental Conditions.

Among the pretrained architectures evaluated, InceptionV3 achieved relatively low training accuracy, reaching approximately 40–50%, with corresponding validation accuracy showing notable fluctuations. ResNet50 demonstrated the lowest performance, with both training and validation accuracy remaining below 20% throughout training. MobileNetV2 showed improved performance relative to InceptionV3 and ResNet50, with training accuracy exceeding 60%; however, its validation accuracy remained variable and lower than that of the proposed model.

4.6 Final CVMaXX Automated Classification System

The ROS-based pilot experiments demonstrated the importance of achieving balanced class representation, indicated that data augmentation was not required, and supported the adoption of a two-stage deep convolutional neural network framework involving sequential vertebral identification and CVM stage classification. These findings informed the architectural and training refinements incorporated into the final CVMaXX system, which is presented in the subsequent section.

The final dataset comprised 1,047 digital lateral cephalometric radiographs from patients aged 6–17 years (mean age 12.4 ± 2.8 years**).** Age distribution across CVM stages demonstrated a progressive increase in mean age with advancing maturation stage, consistent with expected growth trajectories (Table 4.5). In terms of stage representation, CVM Stages III to VI were intentionally balanced for model development, each comprising 19.1% of the dataset, while CVM Stage II accounted for 14.5%, and CVM Stage I represented the smallest proportion at 9.1% (Tables 4.5; Figure 4.11).

Sex distribution across CVM stages is summarised in Table 4.6. Early CVM stages showed a higher representation of male subjects, whereas mid-pubertal stages (CVM Stages III and IV) demonstrated relatively balanced sex distribution compared with other CVM stages. In contrast, later maturation stages (CVM Stages V and VI) were characterised by a predominance of female subjects. Overall, the dataset comprised 457 male (43.6%) and 590 female (56.4%) subjects. Following stage-wise balancing during dataset curation, sex distribution across CVM stages demonstrated a stage-dependent pattern in skeletal maturation timing across sexes.

The ethnic composition of the study sample was predominantly Malay (77.8%), followed by Chinese (16.5%) and Indian (5.7%) subjects (Table 4.8). Within individual CVM stages, Malay participants constituted the majority across all stages, while Chinese and Indian participants were consistently represented, reflecting the inclusion of multiple ethnic groups within the Malaysian clinical sample (Table 4.7).

Table 4.5
Demographic Distribution of the Dataset by CVM Stage, Showing Counts, Percentages, Mean Age, and Age Ranges.

CVM Stage	Count	Percentage %	Mean Age (years)	Age Range
I	95	9.1	7.8± 1.2	6.1-10.2
II	152	14.5	9.2 ± 1.5	6.8-12.1
III	200	19.1	10.9 ± 1.4	8.2-13.5
IV	200	19.1	12.7 ± 1.6	9.8-15.2
V	200	19.1	14.1 ± 1.8	11.2-16.8
VI	200	19.1	15.9 ± 1.2	13.5-17.0
Total	1047	100	12.4 ± 2.8	6.1-17.0

Table 4.6
Sex Distribution across CVM Stages

CVM Stage	Male, n (%)	Female, n (%)	Total, n (%)
I	51 (53.7%)	44 (46.3%)	95 (100%)
II	96 (63.2%)	56 (36.8%)	152 (100%)
III	103 (51.5%)	97 (48.5%)	200 (100%)
IV	103 (51.5%)	97 (48.5%)	200 (100%)
V	45 (22.5%)	155 (77.5%)	200 (100%)
VI	59 (29.5%)	141 (70.5%)	200 (100%)
Total	457 (43.6%)	590 (56.4%)	1047 (100%)

Table 4.7
Ethnic Composition within each CVM Stage. Percentages are Calculated within each CVM Stage.

CVM Stage	Malay, n (%)	Chinese, n (%)	Indian, n (%)	Total (n)
I	73 (76.8)	11 (11.6)	11 (11.6)	95
II	119 (78.3)	22 (14.5)	11 (7.2)	152
III	162 (81.0)	29 (14.5)	9 (4.5)	200
IV	132 (66.0)	49 (24.5)	19 (9.5)	200
V	146 (73.0)	46 (23.0)	8 (4.0)	200
VI	182 (91.0)	16 (8.0)	2 (1.0)	200

Table 4.8
Overall Ethnic Composition of the Study Sample. Percentages Are Calculated based on the Total Study Sample (N = 1,047).

Ethnicity	n	%
Malay	814	77.8%
Chinese	173	16.5%
Indian	60	5.7%
Total	1,047	100%

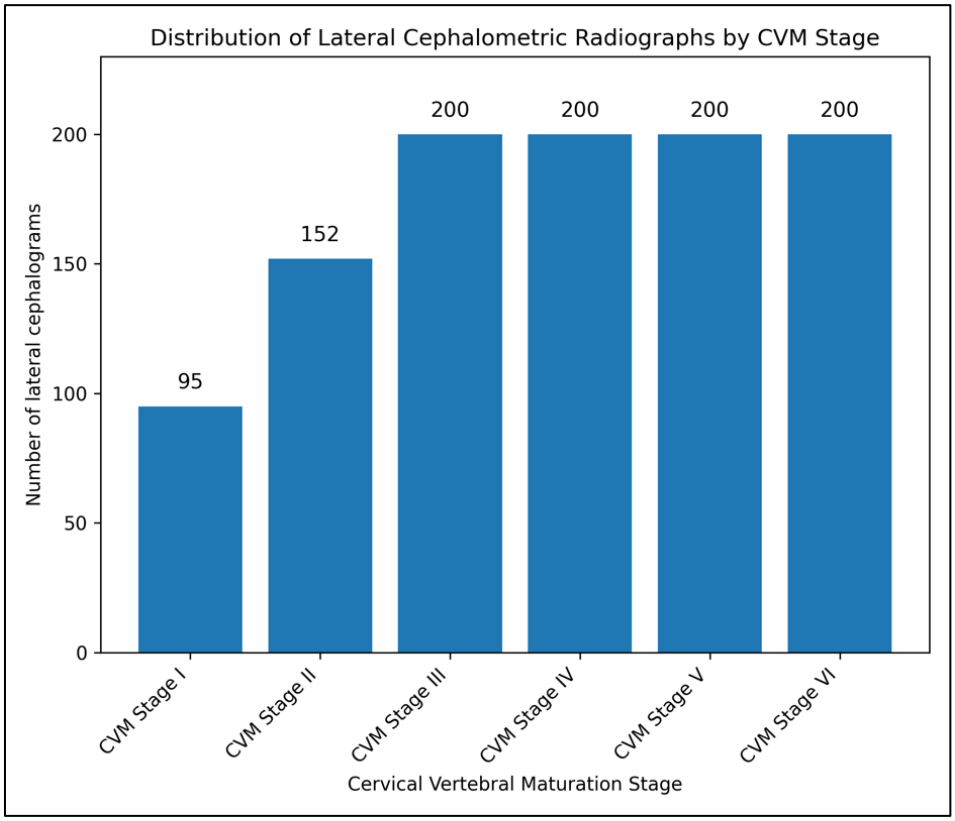


Figure 4.11 Distribution of Lateral Cephalometric Radiographs across CVM Stages.

The reliability analysis demonstrated excellent agreement across intra-examiner, inter-examiner, and AI-versus-consensus assessments for CVM staging (Table 4.9). Intra-examiner reliability was very high, with ICC of 0.967 (95% CI: 0.958–0.974) for Examiner 1 and 0.964 (95% CI: 0.954–0.972) for Examiner 2. Inter-examiner reliability was similarly robust, with an ICC of 0.946 (95% CI: 0.930–0.959) and a weighted Cohen's kappa of 0.852 (95% CI: 0.831–0.873).

Table 4.9
Intra- and Inter-Examiner Reliability Metrics for CVM Staging.

Measure	Examiner 1	Examiner 2	Inter-examiner
ICC (95% CI)	0.967 (0.958-0.974)	0.964 (0.954-0.972)	0.946 (0.930-0.959)
Weighted Kappa	0.891 (0.875-0.907)	0.887 (0.870-0.904)	0.852 (0.831-0.873)
Pearson r	0.923	0.919	0.898
P-value	<0.001	<0.001	<0.001

4.6.1 Stage 1 (Object detection)

Following successful annotation, the YOLOv8 object-detection achieved a mean Average Precision (mAP@0.5) of 0.987, demonstrating high localisation accuracy, and consistently identified and localized cervical vertebrae C2–C4 with 99% performance across all evaluation metrics and minimal variability between vertebral levels (Figure 4.12). Visual inspection confirmed that the detected bounding boxes consistently enclosed the vertebral regions relevant to CVM staging, demonstrating robustness in varied image qualities and anatomical presentations (Figure 4.13 and 4.14).

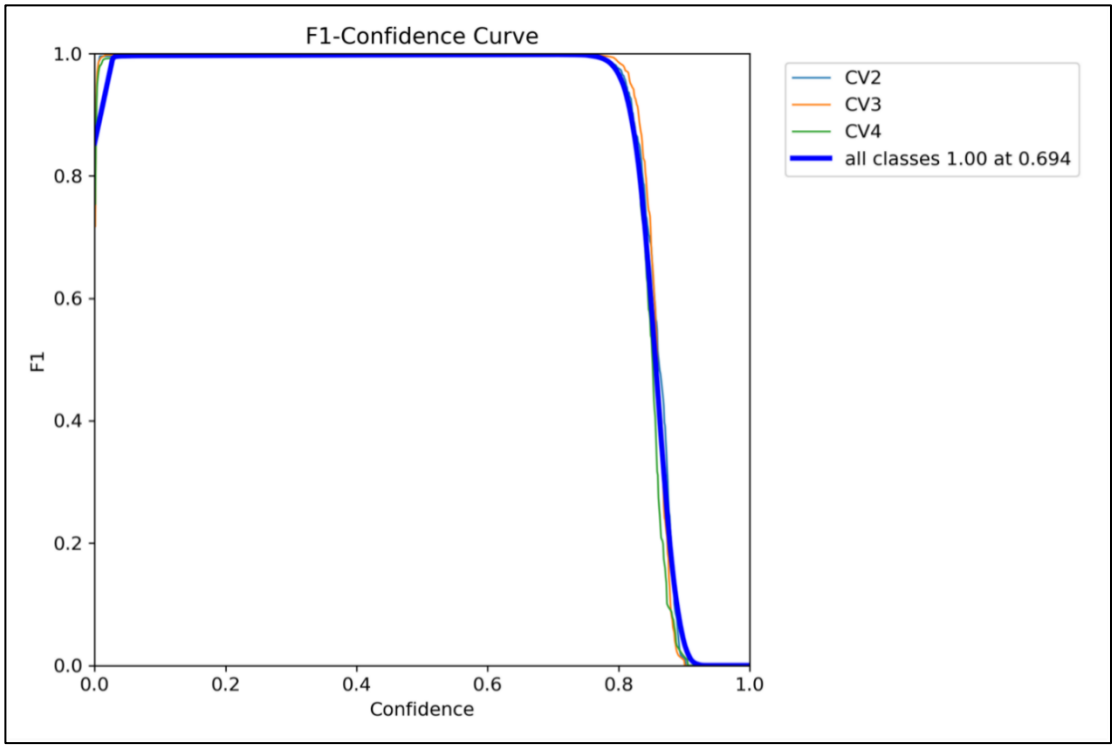


Figure 4.12 F1–Confidence Curve of the YOLOv8 Object Detection Model.

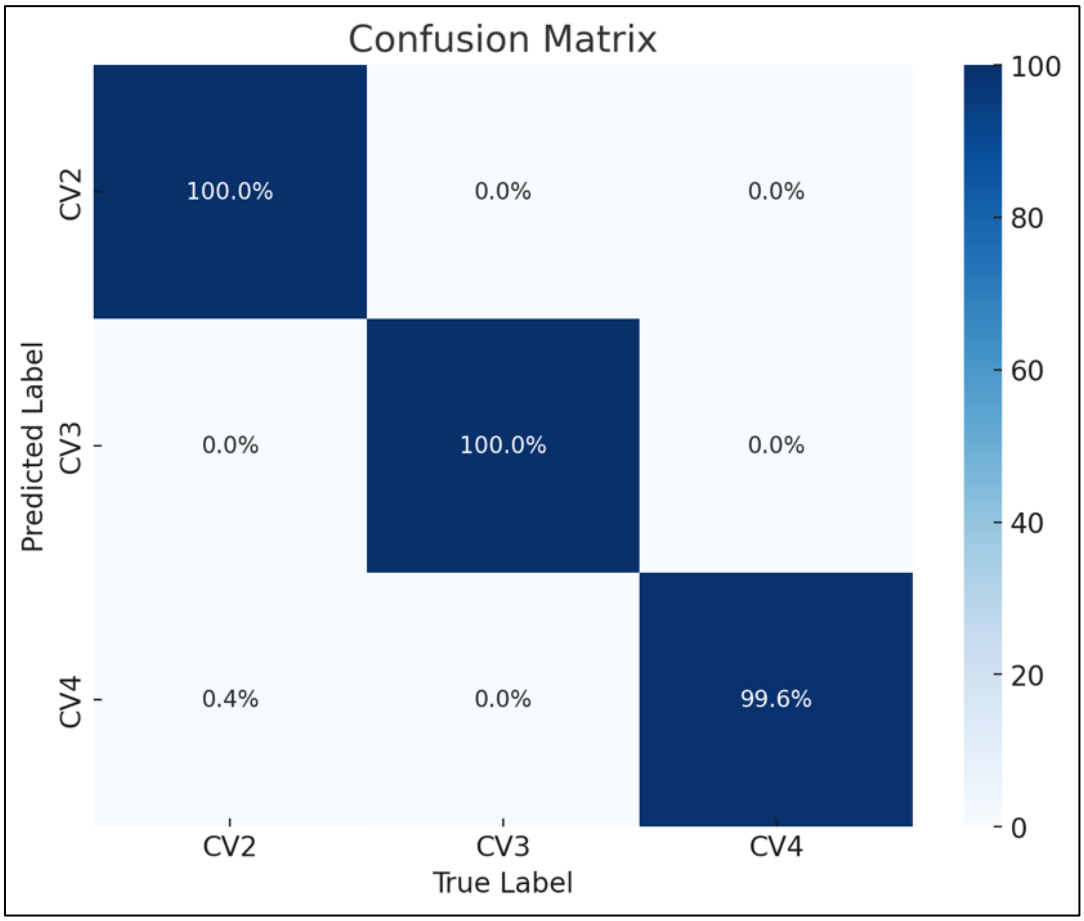


Figure 4.13 Confusion Matrix of the YOLOv8 Object Detection Model.

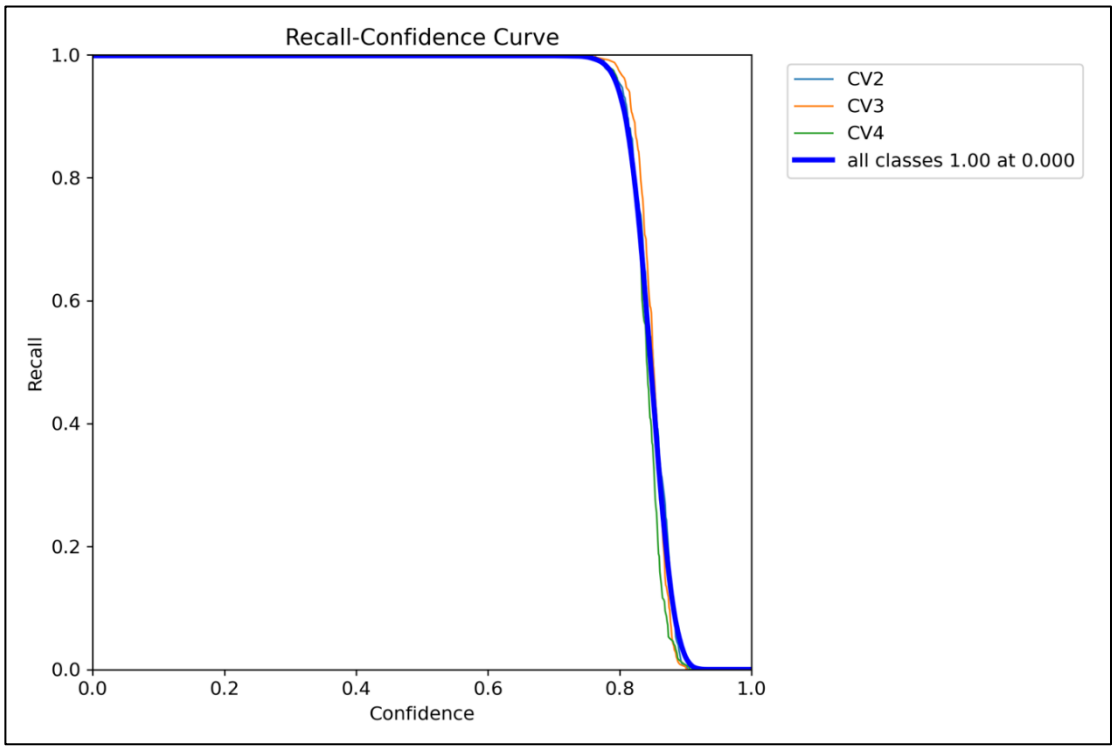


Figure 4.14 Recall–Confidence Curve of the YOLOv8 Object Detection Model.

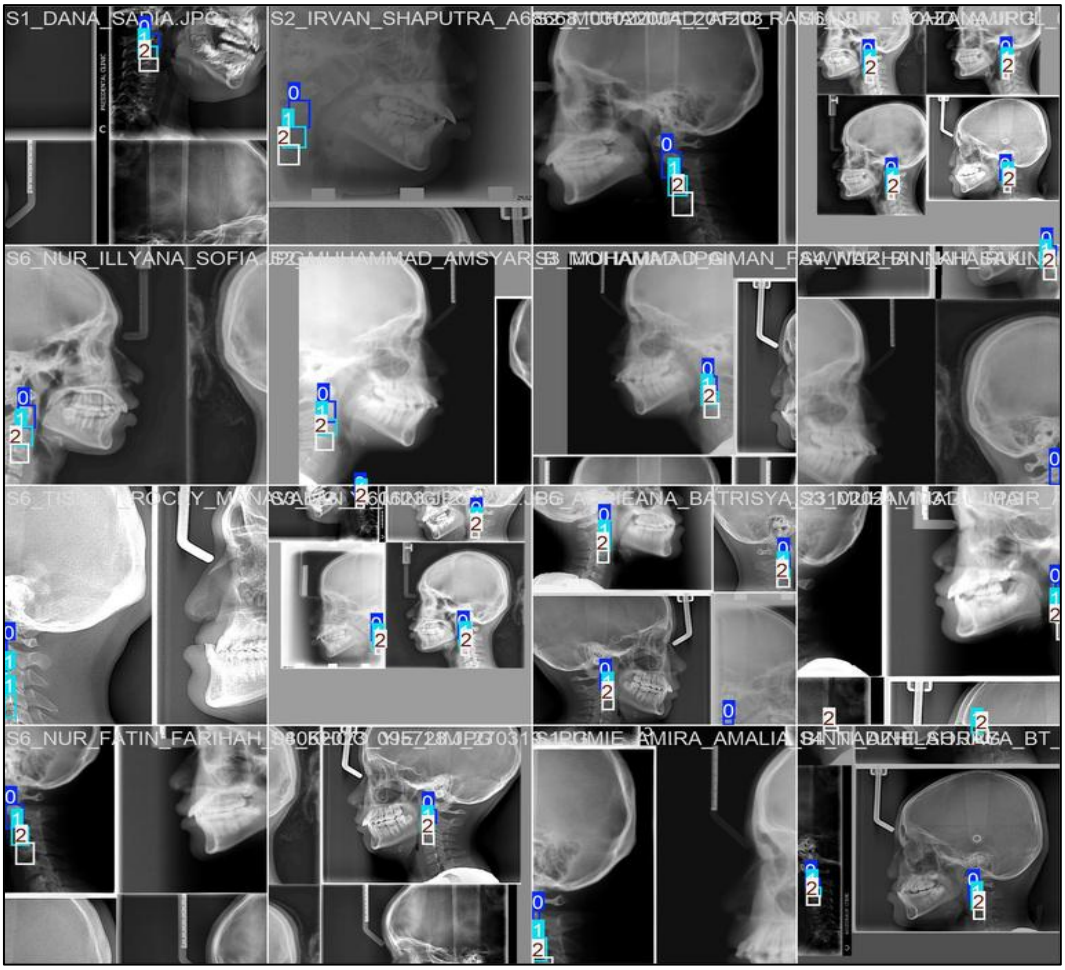


Figure 4.15 Representative Outputs of Automated C2–C4 Vertebral Localization.

Example outputs from the CVMaXX framework showing automated localisation of vertebral bodies with bounding boxes. Detected regions are labelled with class identifiers (“0”, “1”, “2”), corresponding to C2, C3, and C4, demonstrating accurate segmentation of cervical vertebrae across diverse lateral cephalometric radiographs. (Figure 4.15). The cropped vertebral images were then analyzed by each classifier for stage prediction (Figure 4.16).

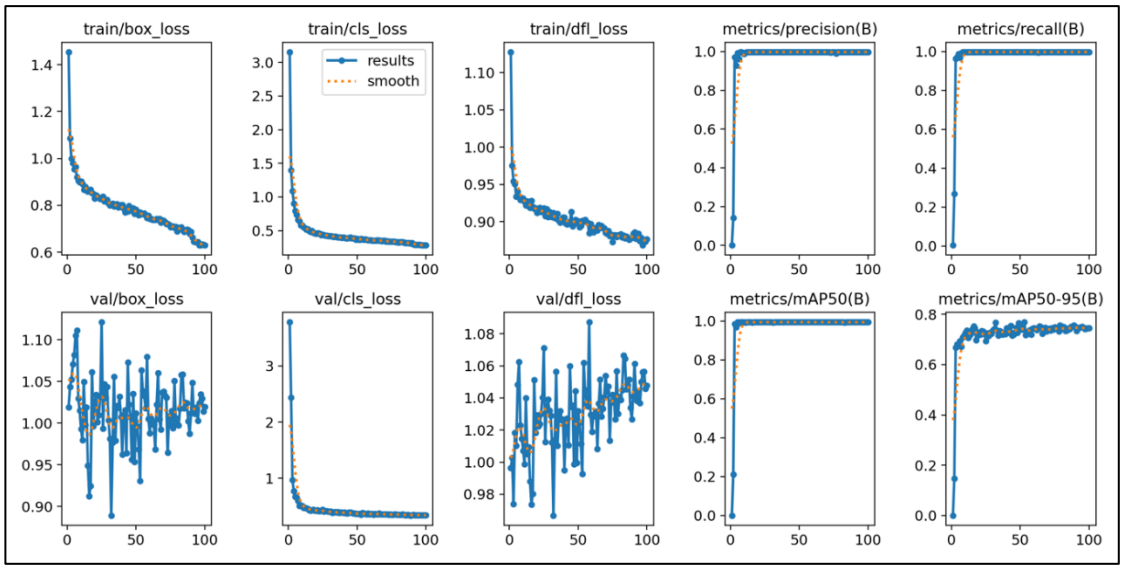


Figure 4.16 Training and Validation Metrics of the YOLOv8 Vertebral Localisation Model.

4.6.1.1 Classification Accuracy and Stage-wise Performance

The confusion matrix illustrates the model's performance across all six CVM stages on the training dataset, yielding an overall accuracy of 91%. Misclassifications were primarily observed in the transitional stages, with CVM Stage II (acceleration phase) achieving 66.7% accuracy due to 16.7% confusion with both CVM Stage I and CVM Stage IV, and CVM Stage IV (deceleration phase) reaching 80.0% accuracy with 20.0% misclassification as CVM Stage III (Figure 4.17). In contrast, perfect classification accuracy (100%) was achieved for four morphologically distinct stages: CVM Stage I (initiation), CVM Stage III (transition), CVM Stage V (maturation), and CVM Stage VI (completion). Overall, the model demonstrated strong performance with an average accuracy of 91% across all six CVM stages (Figure 4.17).

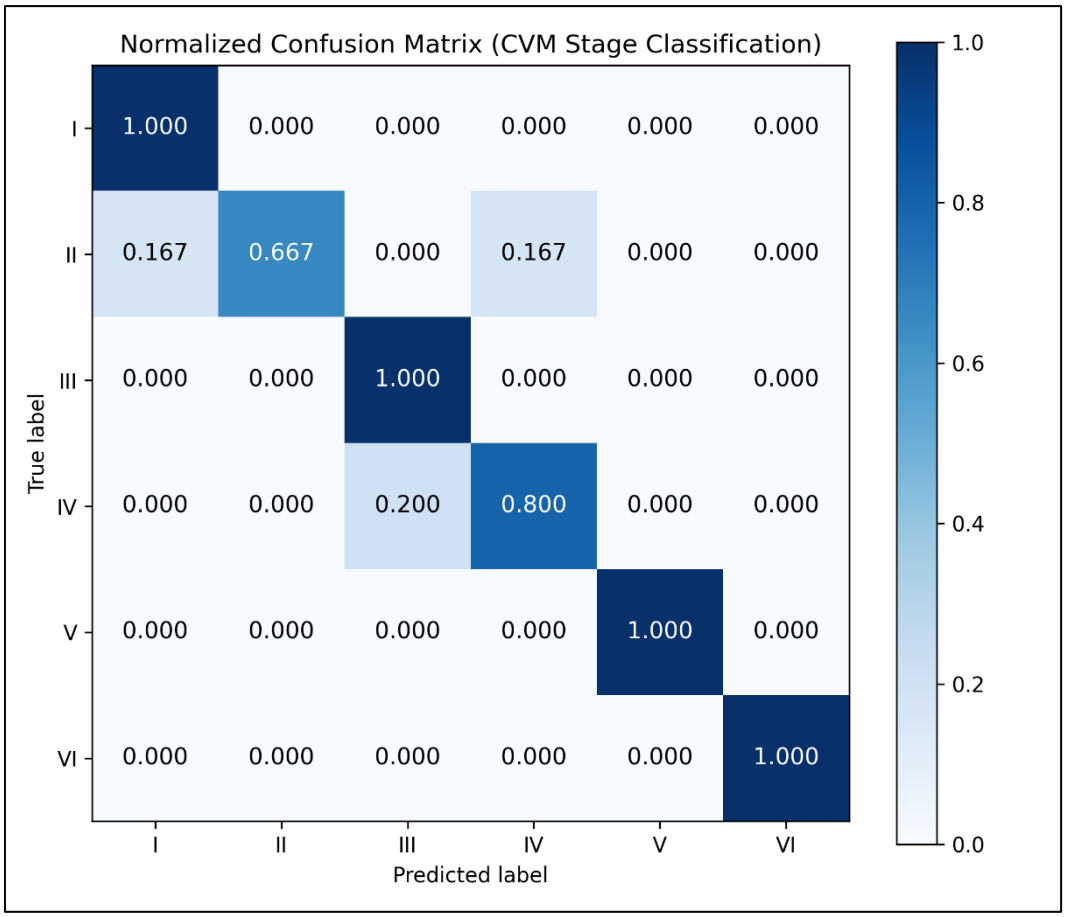


Figure 4.17 Normalised Confusion Matrix of CVM Stage Classification on the Training Dataset.

4.6.2 Stage 2 (Morphological Analysis and Classifier Selection)

The comparative analysis of three machine learning classifiers revealed distinct performance characteristics in CVM classification. LightGBM achieved the highest Area Under the Curve (AUC) value of 0.98, followed by XGBoost (AUC = 0.94) and Random Forest (AUC = 0.90) (Figure 4.18). The Receiver Operating Characteristic (ROC) curve for LightGBM demonstrated rapid ascent toward the upper-left corner indicates near-perfect sensitivity and specificity, while XGBoost also showed strong discrimination with its curve closely following the upper-left ROC boundary. In contrast, Random Forest exhibited a more gradual curve ascent, reflecting comparatively lower sensitivity at lower false positive thresholds. The diagonal line represents chance-level classification. The differences in classifier performance were statistically significant ($p < 0.001$, DeLong test for AUC comparison). All three

classifiers exceeded the threshold for excellent diagnostic accuracy (AUC > 0.90), confirming their clinical utility for automated CVM assessment.

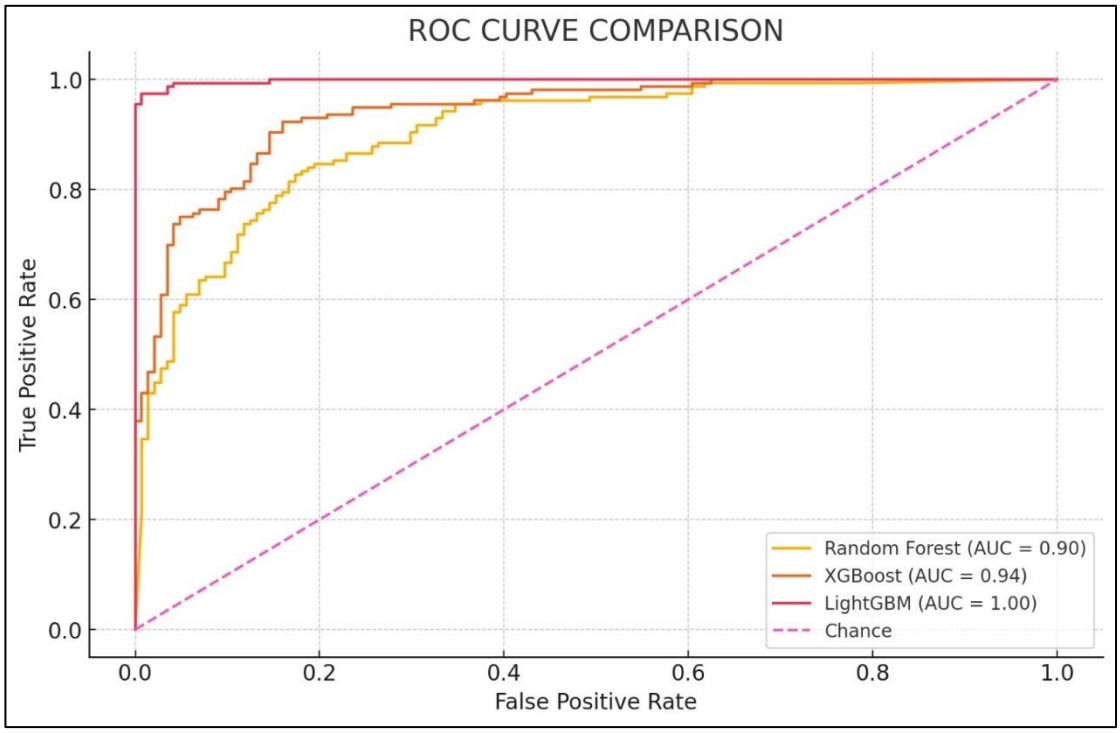


Figure 4.18 ROC Curve Comparison of Three Classifiers for CVM Stage Classification.

4.7 Phase 3: Performance Evaluation of CVMaXX

The performance of the CVMaXX framework was evaluated using internal testing and external validation datasets to quantify classification accuracy, stage-wise behaviour, and agreement with expert consensus. Using morphology-based features extracted from YOLOv8-localised vertebral regions, the LightGBM classifier achieved an overall classification accuracy of 96% for CVM stage prediction. Stage-wise performance demonstrated high accuracy for CVM Stage II (97.7%), Stage IV (95.0%), Stage V (100%), and Stage VI (99.0%). CVM Stage III achieved an accuracy of 91.3%, with 8.7% of cases misclassified as CVM Stage IV. CVM Stage I recorded the lowest accuracy (75.6%), with misclassifications occurring predominantly as CVM Stage III (17.8%). Misclassifications across all stages were primarily confined to adjacent CVM stages. The distribution of predicted versus true CVM stages is illustrated in the normalised confusion matrix (Figure 4.19), with representative classification probability outputs shown in Figure 4.20.

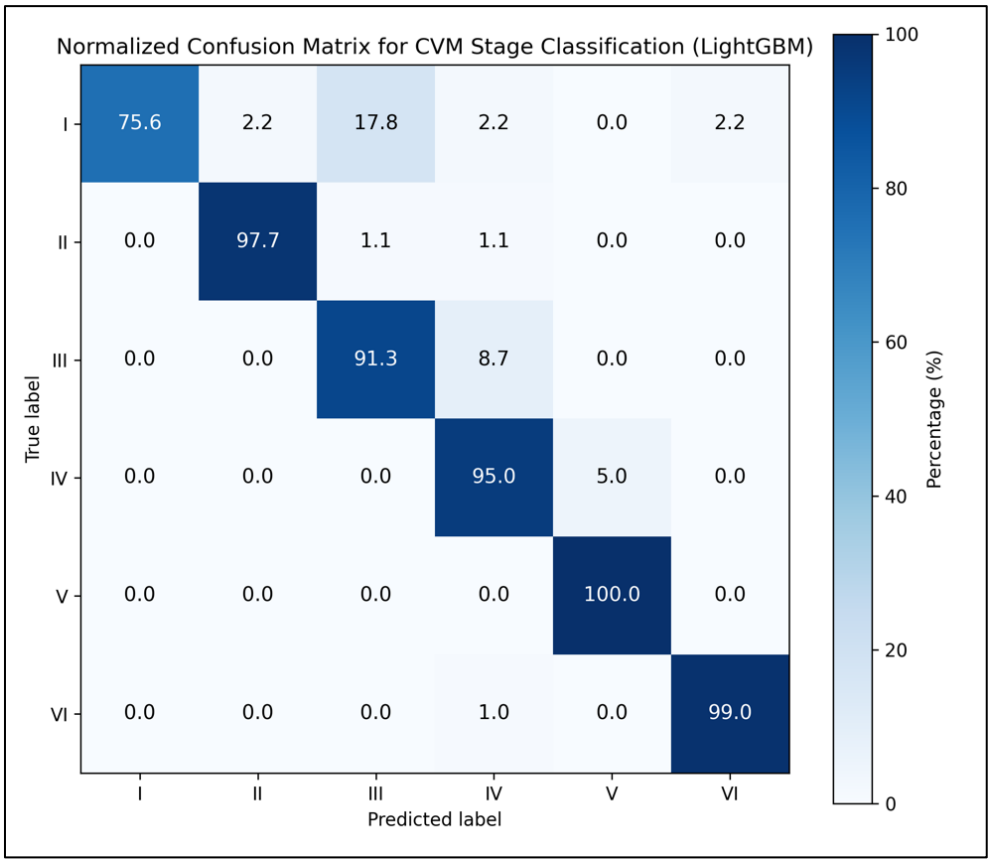


Figure 4.19 Normalised Confusion Matrix of the LightGBM Classifier for Morphological CVM Stage Analysis.

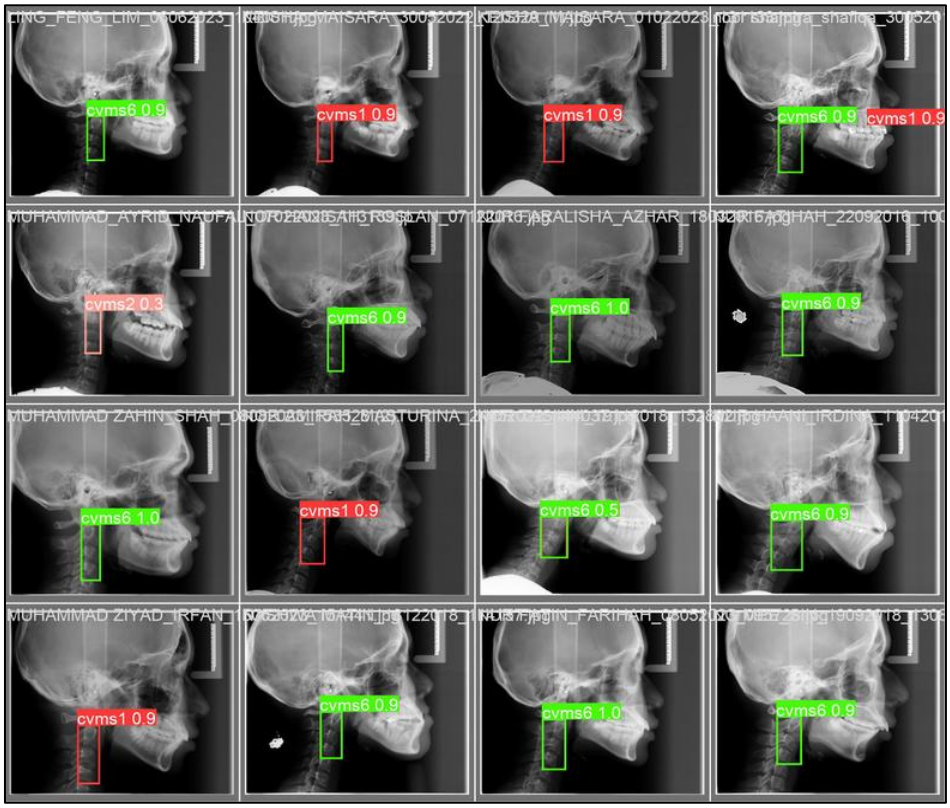


Figure 4.20 Representative Outputs of LightGBM-based CVM Stage Classification with Predicted Probabilities.

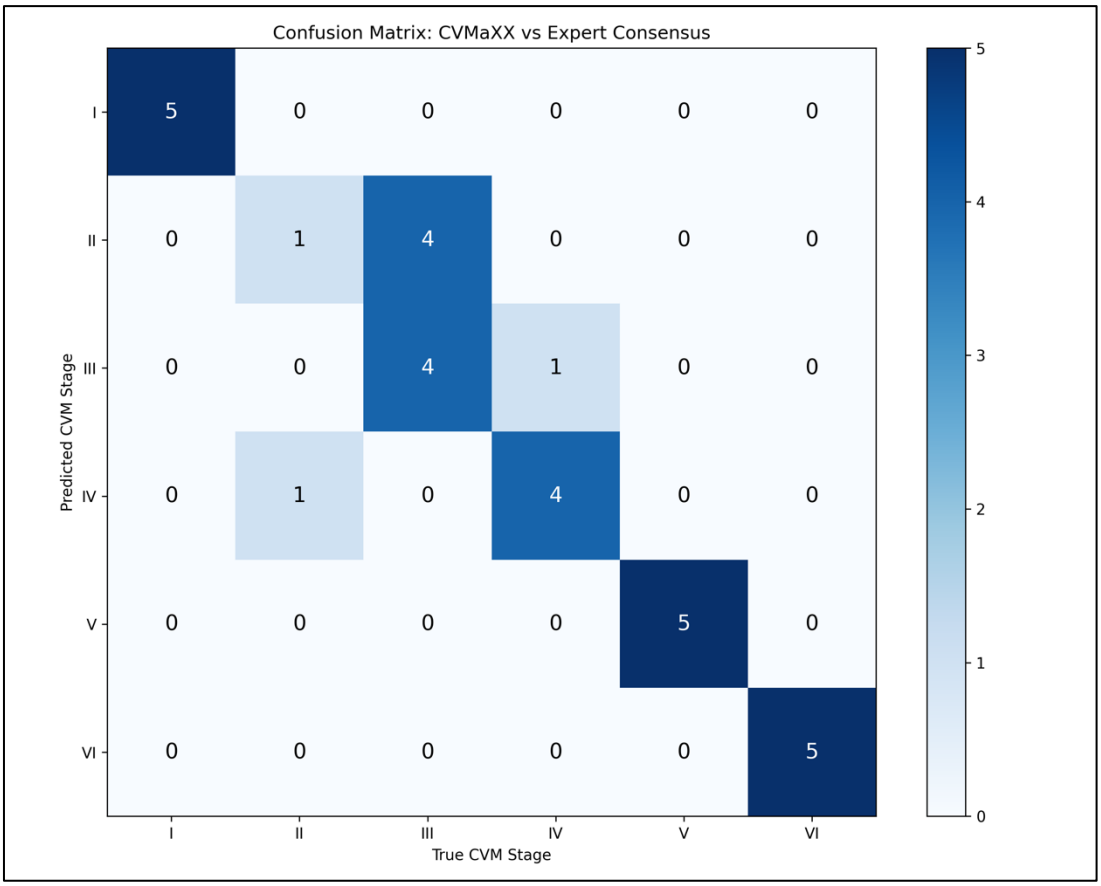


Figure 4.21 Confusion Matrix Comparing CVMaXX Predictions with Expert Consensus on the Independent Unseen Test Dataset.

External validation of the proposed two-stage CVMaXX framework was conducted using an independent unseen dataset comprising 30 lateral cephalometric radiographs. Model performance was assessed by comparing predicted CVM stages with expert consensus annotations. Overall, 24 cases demonstrated exact stage agreement, while six cases showed disagreement, yielding a strict accuracy (exact CVM stage agreement) of 80.0%. Importantly, no misclassifications exceeded one CVM stage, indicating that all predictions fell within clinically acceptable limits. Model performance was evaluated using inclusive and strict evaluation criteria, as defined in Section 3.6.3, to reflect both clinically tolerant and conservative assessment frameworks. Performance outcomes from the external validation are summarised in Table 4.10 and Figure 4.21.

Table 4.10
Performance of the CVMaXX Framework under Inclusive and Strict Evaluation Criteria.

Metric	Value
True Positives (TP)	24
True Negatives (TN)	5
False Positives (FP)	0
False Negatives (FN)	1
Accuracy (Overall evaluation)	96.67%
Accuracy (TP-Only evaluation)	80.00%

Inclusive Evaluation (Exact and Adjacent Stage Agreement)

Using the inclusive evaluation criterion, which considered predictions within ±1 CVM stage as correct, the model achieved an overall accuracy of 96.7%. This result indicates highly consistent performance across unseen radiographs when allowance was made for clinically acceptable transitional variation between adjacent CVM stages. Misclassifications were infrequent, resulting in minimal disagreement beyond one stage.

Strict Evaluation (Exact Stage Agreement Only)

Under strict evaluation criteria requiring exact stage concordance, the CVMaXX framework achieved an accuracy of 80.0% on the independent unseen dataset. While this metric reflects conservative classification correctness, it does not fully capture the clinical plausibility of model predictions. Importantly, all misclassifications were confined to adjacent CVM stages, consistent with the known biological continuity of skeletal maturation. No predictions exceeded one-stage disagreement, indicating that errors remained clinically bounded and unlikely to influence treatment timing decisions.

Agreement between CVMaXX predictions and expert consensus was therefore additionally assessed using Cohen's unweighted kappa (κ), which provides a conservative estimate of case-level agreement beyond chance. The resulting κ value of approximately 0.76 indicates substantial agreement, comparable to that reported among

calibrated human examiners (Figure 4.22). This measure treats all disagreements equally and was selected to prioritise reliability and diagnostic correctness in early system validation. Supplementary weighted kappa analysis, reported in Chapter 5, further demonstrates that most discrepancies represented minor adjacent-stage differences, reinforcing the clinical acceptability of the automated staging outcomes.

Formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Observed agreement:

$$P_o = \frac{\sum n_{ii}}{N} = \frac{96}{120} = 0.80$$

$$\frac{24}{30} = 0.80$$

Expected agreement (Pe):

$$P_e \approx \sum_{i=1}^{6} (0.167 \times 0.167) \approx 0.167$$

Result:

$$\kappa = \frac{0.80 - 0.17}{1 - 0.17} \approx 0.76$$

Figure 4.22 Calculation of Cohen's kappa for Agreement between AI-based CVM Staging and Examiner Consensus.

While unweighted Cohen's κ indicated substantial agreement (κ = 0.76), linear weighted κ increased to 0.89 and quadratic weighted κ reached 0.95, reflecting near-

perfect agreement when accounting for the ordinal structure of CVM stages (Figure 4.23).

Linear weighted κ, $w_{ij}^{\text{linear}} = 1 - \frac{|i-j|}{k-1}$ ≈ 0.89

Quadratic weighted κ, $w_{ij}^{\text{quadratic}} = 1 - \left(\frac{i-j}{k-1}\right)^2$ ≈ 0.95

k = number of categories (6 CVM stages)

i, j = true and predicted categories (1 to 6)

Figure 4.23 Calculation of Weighted Cohen's kappa for Agreement between AI-based CVM Staging and Examiner Consensus.

Agreement between CVMaXX predictions and examiner consensus was substantial using unweighted Cohen's κ (0.76; 95% CI: 0.57–0.92) with overall accuracy of 80%, under strict evaluation criteria, 96.7% inclusive accuracy (allowing ±1 CVM stage agreement) and increased to almost perfect with linear weighted κ (0.89) and near perfect with quadratic weighted κ (0.95), indicating that most disagreements were minor and confined to adjacent CVM stages.

4.8 Deployed System Architecture and Online Workflow of CVMaXX

The overall workflow of the CVMaXX automated classification system is presented in Figure 4.24. The process was structured into four sequential stages to ensure systematic and reproducible analysis. The first stage involved data collection, where 1,047 digital lateral cephalometric radiographs were curated, pre-processed, and partitioned into training, validation, and testing subsets. The second stage employed object detection using YOLOv8 to automatically localize cervical vertebrae C2–C4 with high accuracy. In the third stage, morphological analysis was performed, extracting 72 optimized geometric and textural features from the localized vertebral regions. Finally, in the fourth stage, these features were processed by the LightGBM classifier,

which assigned each case to one of six CVM stages. Together, these stages formed an integrated two-step pipeline that enabled fully automated skeletal maturity assessment.

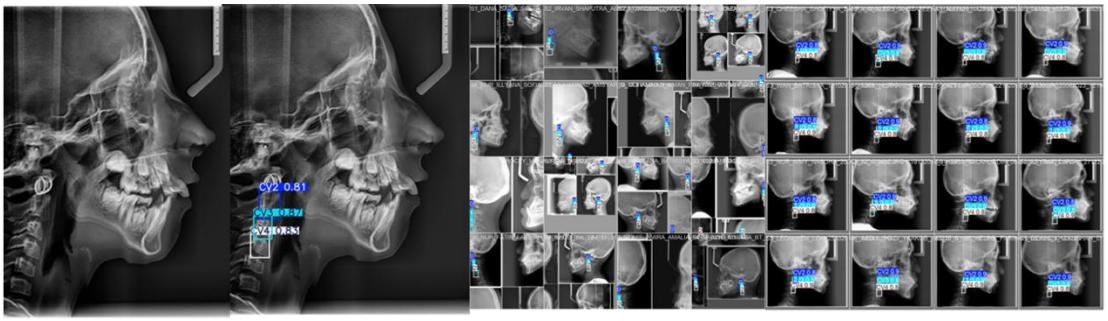


Figure 4.24 Workflow Outputs of the Automated CVM Classification System, Illustrating the Four Main Stages: Data Collection, Object Detection, Morphological Analysis, and Classification.

Figure 4.24 illustrates the proposed CVMaXX framework following full integration of the front-end user interface and back-end AI pipeline. The system represents an end-to-end automated workflow, beginning with digital lateral cephalometric image input via a web-based interface and progressing through data storage, preprocessing, object detection, morphological feature extraction, and CVM stage classification. The back-end comprises a two-stage deep learning architecture incorporating vertebral localisation and feature-based classification, while the front-end enables case management, visualisation, and retrieval of diagnostic outputs. This integrated design allows seamless interaction between data acquisition, model inference, performance evaluation, and result presentation, supporting both clinical usability and continuous model refinement within the CVMaXX platform. The framework below demonstrates integration of front-end user interaction with back-end AI modules for automated CVM assessment (Figure 4.25).

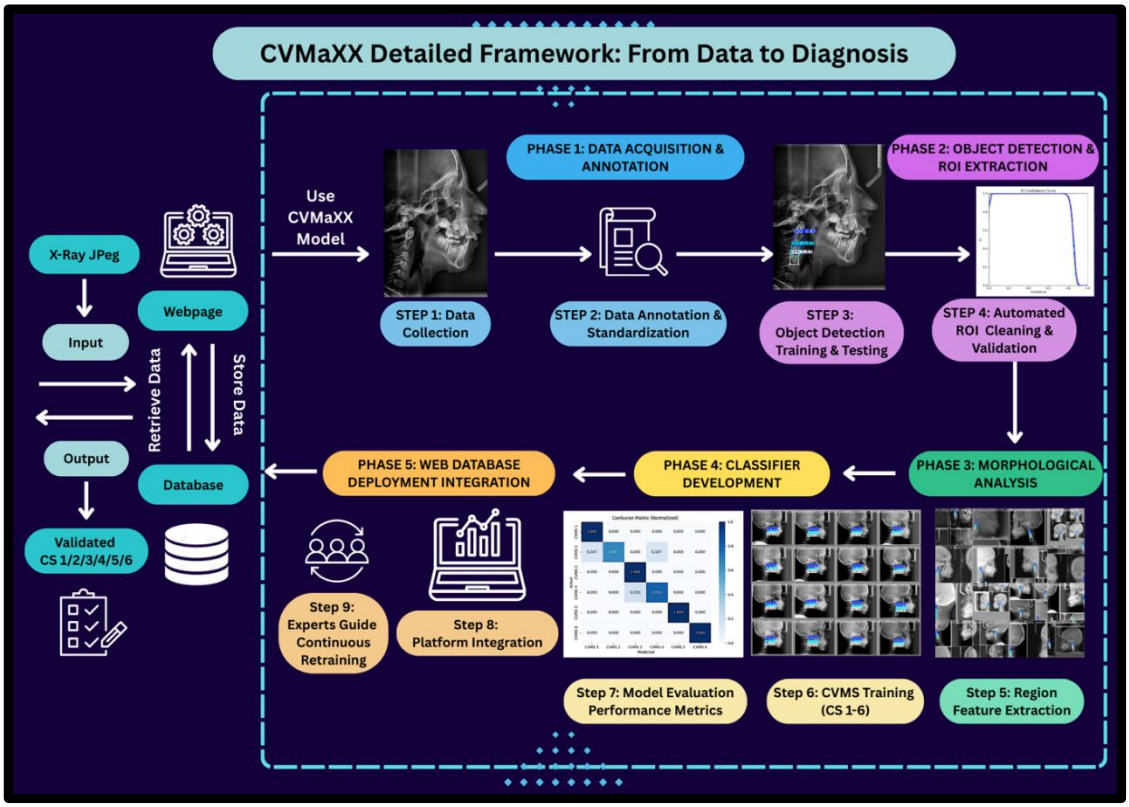


Figure 4.25 Proposed CVMaXX Integrated Framework from Data Input to Automated CVM Diagnosis.

Schematic representation of the fully integrated CVMaXX system, illustrating the end-to-end workflow from lateral cephalometric image acquisition and annotation through object detection, region-of-interest extraction, morphological feature analysis, CVM stage classification, performance evaluation, and web-based deployment (Figure 4.26a-d).

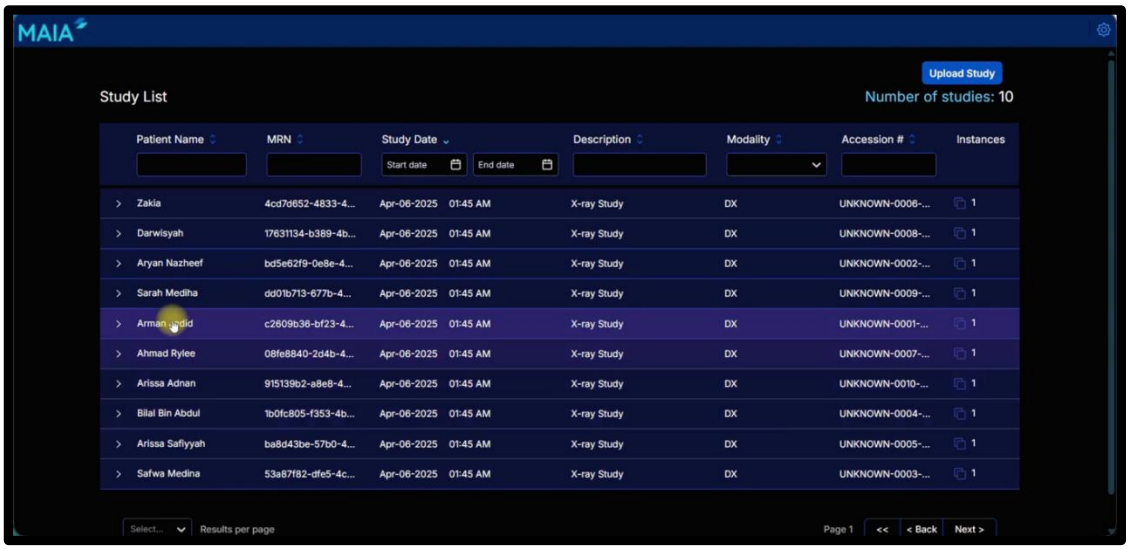


Figure 4.26a CVMaXX Study List Interface.

The CVMaXX system dashboard displaying the list of uploaded lateral cephalometric radiograph studies, including patient identifiers, study date, modality, accession number, and number of image instances. This interface enables structured case management prior to automated analysis.

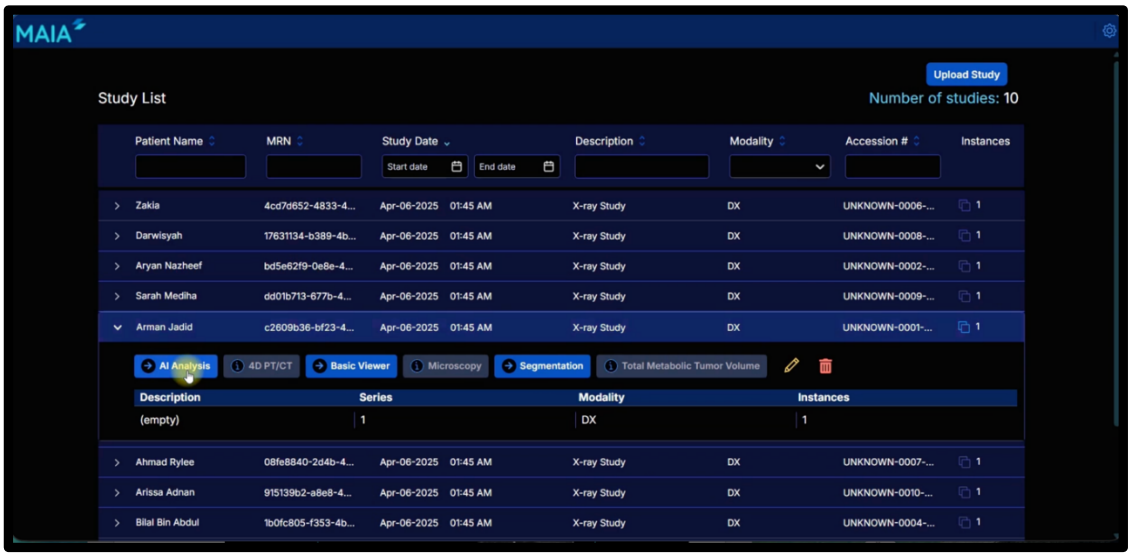


Figure 4.26b CVMaXX Study Selection and AI Analysis Initiation.

Expanded study view illustrating case selection and access to AI-assisted functions, including initiation of automated CVM analysis. The interface allows users to review study metadata and trigger downstream processing within the CVMaXX pipeline.

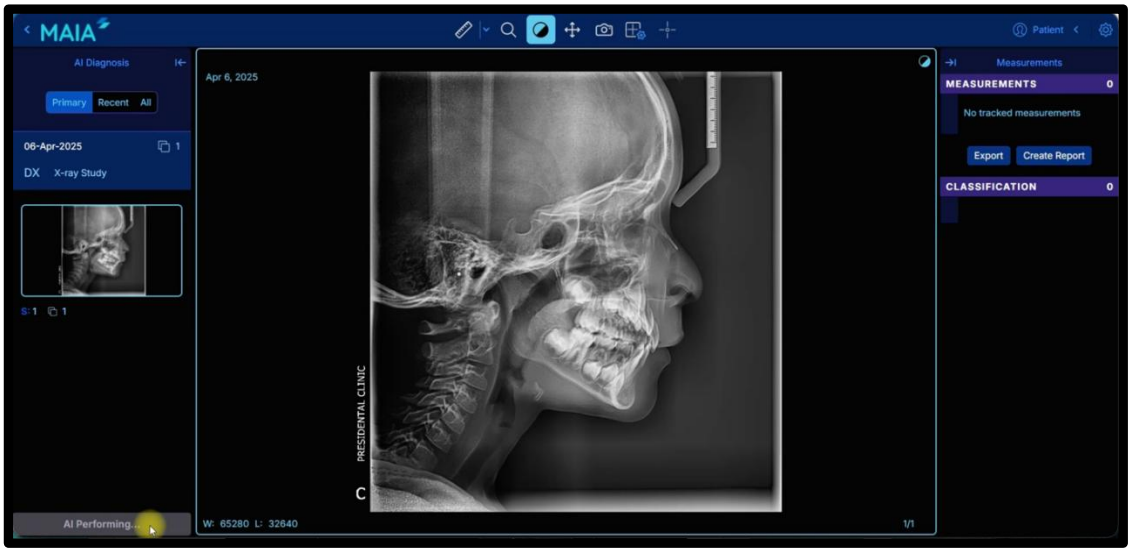


Figure 4.26c Lateral Cephalometric Radiograph Visualisation within CVMaXX.

Display of a lateral cephalometric radiograph within the CVMaXX viewer prior to automated classification. The image is presented with standard viewing controls, serving as the input for cervical vertebral localisation and maturation assessment.

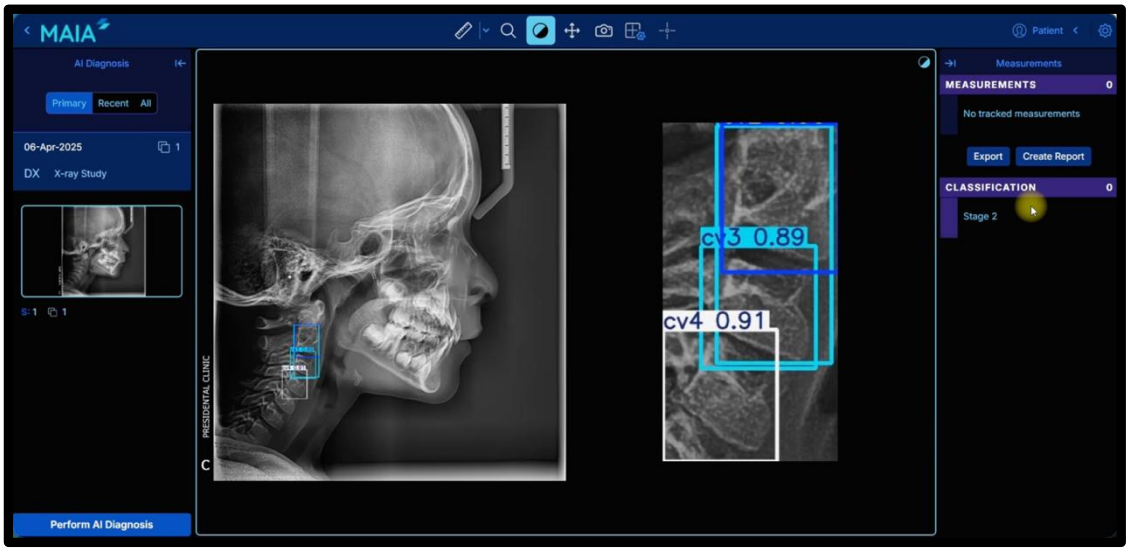


Figure 4.26d Automated Cervical Vertebra Localisation and CVM Stage Output.

Automated detection of cervical vertebrae (C3 and C4) using the CVMaXX object detection module, with bounding boxes and confidence scores overlaid. The corresponding predicted CVM stage is displayed in the classification panel, demonstrating end-to-end automated CVM assessment.

CHAPTER 5
DISCUSSION

5.1 Introduction

This chapter interprets the findings of the present study in relation to its research objectives and research questions, situating the results within the broader orthodontic and medical artificial intelligence literature. The discussion follows the sequential development of the CVMaXX framework, beginning with examiner reliability and ground truth dataset construction, progressing through model development and performance evaluation, and concluding with considerations of clinical relevance.

5.2 Phase 1: Dataset Development and Pilot Study

5.2.1 Pilot Feasibility Study

Before assembling the main dataset, a pilot feasibility study using 20 archived lateral cephalometric radiographs was conducted to determine whether Baccetti's CVM staging method could be reliably applied to Malaysian population. The aim was to assess stage interpretability, transitional-stage ambiguity, and examiner variability in a local clinical context. Transitional stages exhibited overlapping morphological features, consistent with the biological progression of pubertal growth, where skeletal maturation unfolds gradually rather than through discrete structural shifts. This continuity complicates categorical staging, particularly at the acceleration and deceleration phases of puberty, and has been associated with systematic reductions in staging agreement among expert raters (Perinetti et al., 2018; Schoretsaniti et al., 2021; Zhao et al., 2012).

The pilot confirmed the need for broader representation of transitional and early developmental stages in the final dataset to support both examiner calibration and machine learning analyses. All pilot images were excluded from subsequent modelling to avoid contamination of ground truth annotations, ensuring that feasibility insight informed dataset design without influencing model evaluation, consistent with established practices in medical imaging research (Alhamady et al., 2024; Esteva et al., 2021; Litjens et al., 2017).

5.2.2 Sample Size Determination

Determining the dataset size required balancing two distinct methodological demands: the statistical requirements for multi-examiner reliability analysis and the empirical needs of supervised machine learning (Atici et al., 2022; Jiang et al., 2025). Reliability studies generally require modest sample sizes to estimate inter- and intra-observer agreement with sufficient power whereas deep learning architectures benefit from substantially larger and more morphologically diverse datasets to avoid overfitting and support stable optimisation (Litjens et al., 2017; Schoretsaniti et al., 2021). Given that transitional stages in CVM staging are both clinically important and prone to diagnostic uncertainty, the dataset was intentionally expanded beyond the minimum reliability threshold to improve representation across the full growth continuum. This approach aligns with recommendations in orthodontic and medical AI research, where class balance and morphological diversity contribute more meaningfully to model generalisability than raw sample size alone (Atici et al., 2022; Buda et al., 2018; Chawla et al., 2002; Johnson & Khoshgoftaar, 2019). The resulting dataset therefore supported both reliability assessment and experimental modelling, establishing a methodologically stronger foundation for subsequent development of the CVMaXX system.

5.2.3 Dataset Characteristics

The final dataset comprised lateral cephalometric radiographs from Malaysian children and adolescents aged 6 to 17 years (mean age 12.4 ± 2.8 years), capturing the developmental period spanning from prepubertal to postpubertal skeletal maturation. Class representation was intentionally balanced across CVM Stages I to VI to mitigate class imbalance, a well-recognised challenge in supervised classification tasks where uneven distributions lead to inflated performance metrics and reduced sensitivity for clinically significant minority stages (Amasya, Yildirim, et al., 2020; Atici et al., 2022; Buda et al., 2018; He & Garcia, 2009; Johnson & Khoshgoftaar, 2019; Seo et al., 2021). This stage-wise balancing ensured that the automated classification framework was trained on a dataset that reflected the full spectrum of skeletal maturation rather than being dominated by late pubertal stages, as is commonly observed in orthodontic imaging datasets.

Sex distribution across CVM stages demonstrated a stage dependent pattern that aligns with established biological principles of pubertal growth and skeletal maturation. Early CVM stages exhibited a higher representation of male subjects, mid pubertal stages (CVM Stages III and IV) showed near equal distribution between sexes, and later stages (CVM Stages V and VI) were characterised by a predominance of female subjects. This pattern reflects the sexually dimorphic timing of pubertal development, whereby females typically enter and complete skeletal maturation earlier than males (Baccetti et al., 2005; Flores-Mir et al., 2006). While differential health-seeking behaviour may contribute modestly to sex representation within orthodontic samples, the stage specific nature of this distribution rather than a uniform female predominance across all stages supports a predominantly biological explanation rather than simple referral bias (Shaw et al., 1991). Importantly, sex was not used as an explicit input feature within the automated classification framework. However, the observed distribution provides biological context for interpreting CVM stage assignment and reinforces the relevance of stage based rather than age based or sex based growth assessment (Perillo et al., 2010).

Demographic variables extracted from patient records included age, sex, and ethnicity (Malay, Chinese, and Indian), which were analysed descriptively to characterise the study sample across CVM stages. The overall study sample was predominantly Malay, a distribution that is broadly consistent with Malaysia's national ethnic composition (Department of Statistics Malaysia, 2023). Although Chinese and Indian subjects were comparatively fewer, all major ethnic groups were represented, enabling the model to be trained and evaluated within a multi-ethnic dataset. Ethnic differences in craniofacial growth patterns and maturation trajectories have been documented in orthodontic literature and may influence both skeletal morphology and treatment planning (Flores-Mir et al., 2004; Khazaei et al., 2023). The private clinical data included in this study were obtained from selected clinics within the Klang Valley region and were intended to enhance dataset diversity rather than represent the epidemiological distribution of private orthodontic practice across Malaysia. Furthermore, the substantially larger dataset size relative to earlier studies (typically fewer than 900 images) provided a more diverse and contextually relevant foundation for model development within an ethnically heterogeneous population (Khazaei et al., 2023; Mathew et al., 2022). Expanding the inclusion of minority-group samples in

future studies may further enhance the generalisability and clinical applicability of the proposed framework, particularly for nationwide or regional deployment.

Although stage-wise capping was applied during dataset curation, the availability of early developmental stages (CVM Stages I and II) was inherently constrained by ethical limitations associated with paediatric radiographic acquisition, resulting in comparatively fewer images for these stages (McNamara & Franchi, 2018; Sohrabi et al., 2016). Such constraints are well-recognised in orthodontic and paediatric imaging research, where radiographs are acquired strictly on clinical indication to minimise unnecessary radiation exposure (European Commission, 2015). Despite this limitation, the final dataset reflected real-world orthodontic presentation patterns commonly encountered in clinical practice. Of particular importance, the proposed YOLOv8-based framework demonstrated the capacity to learn effectively from smaller early-stage samples without reliance on data augmentation techniques, aligning with recent evidence supporting the robustness of object detection–driven deep learning architectures in limited-data medical imaging scenarios (Jocher, 2023; Tsirtsakis et al., 2025). Nevertheless, future investigations would benefit from broader early-stage data representation, obtained through ethically compliant strategies, to further strengthen future modelling efforts and improve early intervention applicability.

5.2.4 Clinical and Methodological Implications

The dataset development strategy addressed two persistent limitations in CVM research, namely insufficient diversity in stage representation and limited reliability of expert annotations. Inclusion of a Malaysian cohort contributes important geographic and ethnic diversity to the CVM literature, addressing longstanding gaps in non-Western validation (Flores-Mir et al., 2006).

The retrieval of 1,604 radiographs and subsequent selection of 1,047 diagnostically suitable cases yielded a dataset that is larger and more representative than those used in many prior CVM based artificial intelligence studies, where datasets often comprise fewer than 900 images and demonstrate substantial skew toward later maturation stages (Mohammad-Rahimi et al., 2022; Seo et al., 2021). Importantly, oversampling techniques used in subsequent pilot modelling were applied not to generate a deployable model, but to explore how class balance influences learning behaviour under controlled experimental conditions. This distinction recognises that

CVM staging challenges arise from biological ambiguity and dataset structure rather than inherent limitations of automated classifiers. Collectively, these dataset considerations established a methodologically rigorous platform on which to build the CVMaXX framework, improving both clinical relevance and machine learning feasibility.

5.3 Reliability Analysis and Expert Calibration

5.3.1 Importance of Reliability in CVM Staging

CVM staging is inherently interpretive because it relies on visual assessment of morphologic transitions in cervical vertebrae that progress continuously across adolescence. Unlike binary diagnostic tasks, staging requires assigning discrete categories to a biological process that unfolds gradually, particularly across the pubertal growth acceleration and deceleration phases. Transitional morphological overlap between CVM Stages II, III and IV has been repeatedly shown to depress staging reliability, with disagreement clustering around these biologically ambiguous boundaries (Perinetti et al., 2012; Schoretsaniti et al., 2021; Zhao et al., 2012).

This variability is not primarily attributable to examiner inexperience but to the epistemic structure of the task itself, where categorical labels must be imposed on non-discrete developmental phenotypes. From a clinical perspective, unreliable staging complicates growth-related orthodontic treatment decisions and undermines the value of CVM as an auxiliary indicator for timing interventions. From a machine learning perspective, reliability determines the fidelity of ground truth labels, and inconsistency in annotation produces noisy label-feature mappings that propagate into model decision boundaries and reduce external validity (Esteva et al., 2021; Litjens et al., 2017). Reliable staging is therefore a methodological prerequisite for both clinical use and automated inference.

5.3.2 Multi-Expert Calibration Framework

A structured multi expert calibration framework was implemented to reduce interpretive variability and establish a reproducible consensus ground truth for CVM

staging. Calibration focused primarily on transitional stages, where observer disagreement is known to concentrate and where progressive, overlapping morphological changes exert the strongest influence on staging inconsistency (Perinetti et al., 2018; Schoretsaniti et al., 2021). These transitional intervals coincide with periods of accelerated pubertal skeletal development, during which discrete categorisation becomes inherently challenging due to continuous biological change.

By anchoring interpretation to established patterns of CVM, the protocol provided a consistent strategy for resolving these borderline cases. Interpretation was guided by the recognised cranio-caudal progression of CVM, whereby morphological changes typically appear earlier in the upper cervical vertebrae and advance inferiorly with increasing skeletal maturity (Franchi et al., 2000; McNamara & Franchi, 2018; Sohrabi et al., 2016; Uysal et al., 2006). Within this framework, emphasis was placed on comparing concavity depth and vertebral shape progression across adjacent cervical levels, which helped move interpretation away from broad visual impressions and toward more consistent stage assignment at transitional boundaries.

Importantly, calibration relied exclusively on vertebral morphology observable within lateral cephalograms, without recourse to external biological indicators such as dental maturation. Dentition was explicitly excluded from adjudication because mandibular dental development demonstrates inconsistent synchrony with skeletal maturation and exhibits high inter-individual variability, limiting its reliability as a staging indicator in both clinical and research contexts (Cummaudo et al., 2021; Dadgar et al., 2021).

This restriction ensured biological coherence between expert annotation and the anatomical structures targeted by the automated system, thereby preventing construct leakage between skeletal and dental developmental domains. By formalising expert reasoning as a rule governed interpretive process grounded in known patterns of axial skeletal maturation, calibration reframed CVM staging from a subjective visual judgement into a structured diagnostic task. This approach strengthened both annotation reproducibility and methodological alignment with the subsequent morphology driven classification model (Flores-Mir et al., 2006; Perinetti et al., 2012).

5.3.3 Inter-Observer Reliability Outcomes

Following calibration, inter-observer agreement among thirteen orthodontic examiners improved from moderate ($\kappa = 0.580$) to substantial ($\kappa = 0.611$) levels, surpassing agreement estimates commonly reported in CVM reliability studies employing smaller examiner panels (Perinetti et al., 2012; Schoretsaniti et al., 2021; Zhao et al., 2012).

That such improvements were achieved across examiners drawn from multiple institutions suggests that variability in CVM staging is modifiable and not an immutable property of the method (Sohrabi et al., 2016). This challenges a common assumption in the literature that CVM's clinical utility is fundamentally constrained by subjectivity, instead indicating that structured interpretive rules can mitigate transitional-stage ambiguity and reduce disagreement (Hussain et al., 2024). The resulting consensus therefore represents a more defensible reference standard for both clinical interpretation and supervised model training.

5.3.4 Intra-Observer Reliability Outcomes

Temporal reproducibility also improved, with high intraclass correlation coefficients and Cohen's kappa values demonstrating that calibrated examiners could reapply the criteria consistently when reassessing radiographs after a temporal interval (Hussain et al., 2024; Schoretsaniti et al., 2021; Zhao et al., 2012). Strong intra-examiner agreement is particularly relevant for longitudinal clinical monitoring, where staging must be consistent over time to track maturation and inform treatment timing (Hussain et al., 2024).

For supervised learning, internal consistency reduces stochastic annotation noise, enabling the model to learn stable associations between morphology and stage labels (Litjens et al., 2017). Together, the inter- and intra-observer findings reinforce that structured calibration enables CVM staging to satisfy the minimum reliability expectations required for both clinical decision-making and annotation-driven machine learning.

5.3.5 Impact of the Operational Decision Protocol

The introduction of the operational decision protocol reduced transitional ambiguity not by enforcing uniform interpretations among examiners, but by clarifying how borderline vertebral morphology should be interpreted within a shared biological framework. Ambiguity between adjacent CVM stages is an expected feature of skeletal maturation, as vertebral concavity formation and proportional changes occur gradually rather than discretely, particularly during the pubertal growth spurt when developmental change is rapid and temporally compressed (Baccetti et al., 2005; McNamara & Franchi, 2018). Consequently, early CVM Stage III may be difficult to distinguish from late CVM Stage II, while late CVM Stage III may resemble early CVM Stage IV despite corresponding to different phases of skeletal development.

The calibration process further demonstrated that transitional morphology is frequently heterogeneous rather than uniform. Within both the CVM Stage II to III and Stage III to IV boundaries, early and late transitional presentations were commonly observed, even though the current categorical staging system does not distinguish these sub phases. Clinically, this distinction is relevant because orthodontic treatment timing, particularly for growth modification and interceptive interventions, depends on whether a patient is approaching, experiencing, or exiting the pubertal acceleration phase (Baccetti et al., 2005; Franchi et al., 2000). In this context, transitional ambiguity is not merely diagnostic uncertainty but reflects biological gradation that may warrant more detailed representation in clinical staging and future computational models.

By making transitional cues explicit, the protocol provided a shared reference for interpretation that replaced global visual impressions with structured, stage-specific morphological cues. The improvement in Fleiss' κ observed across calibration rounds suggests that much of the observed variability in unstructured CVM staging reflects uncertainty about how to navigate transitional morphology rather than fundamental disagreement. In this respect, the protocol did not constrain clinical judgement; it merely organised it along a more consistent reading of the biological signal.

Few CVM reliability studies have attempted similar operationalisation. Most acknowledge transitional disagreement as a limitation of CVM without exploring whether it can be structured or reduced (Perinetti et al., 2018; Schoretsaniti et al., 2021). The present findings support an alternative interpretation, namely that transitional ambiguity is intrinsic to continuous skeletal maturation and that explicit operational

handling of this ambiguity enables more consistent application in both clinical assessment and AI driven modelling.

5.3.6 Reliability and Implications for AI Ground Truth Development

The reliability outcomes established in Sections 5.3.1 to 5.3.5 have implications that extend beyond clinical interpretation to the construction of ground truth datasets for AI. In supervised learning, annotation quality directly constrains model performance, as inconsistencies in expert labels introduce noise into the feature–label relationship and limit generalisability (Esteva et al., 2021; Litjens et al., 2017). CVM staging, as an ordinal and morphology-based classification, is particularly sensitive to such effects, given that adjacent stages are separated by subtle and progressive anatomical changes rather than discrete thresholds.

Within this context, Table 5.1 situates the reliability characteristics of the present study against commonly reported benchmarks in the CVM literature (Caldas et al., 2007; Flores-Mir et al., 2006). While agreement metrics in isolation are not novel, the methodological conditions under which they were achieved are critical for AI applicability. Specifically, substantial inter-observer agreement was maintained despite a large, multi-institutional examiner panel, and excellent intra-observer consistency was demonstrated following structured calibration. These conditions approximate real-world clinical heterogeneity more closely than the single-centre, small-panel designs typical of earlier studies, where agreement may be inflated by shared training environments rather than true label stability (Amasya, Cesur, et al., 2020; Atici et al., 2022).

For AI ground truth development, this distinction is fundamental. Training datasets derived from narrowly calibrated or institution-specific interpretations may perform well internally but degrade when deployed across different populations, imaging protocols, or clinical contexts. By contrast, the present dataset was intentionally constructed to absorb observer diversity through calibration and consensus, thereby encoding clinically realistic variability rather than suppressing it. This approach increases the likelihood that learned representations capture biologically meaningful patterns rather than examiner-specific interpretive rules.

Importantly, the findings also indicate that high-quality ground truth cannot be defined by agreement statistics alone. The deliberate integration of explicit decision

rules, consensus adjudication, and multi-expert calibration reflects a shift from implicit expert judgement toward reproducible annotation logic. For AI systems intended to support clinical decision-making, such transparency is essential, as it enables traceability between model predictions and the biological features they are intended to represent.

Taken as a whole, the reliability outcomes and methodological characteristics synthesised in this section support the suitability of the annotated dataset as reference ground truth for AI-based CVM modelling. By demonstrating stable annotation under heterogeneous and clinically realistic conditions, the present study provides a defensible foundation for supervised learning that moves CVM assessment toward a reproducible, technology-assisted framework. This interpretation directly informs the methodological design and validation strategy described in Chapter 3.

Table 5.1
Comparison of Reliability Outcomes from the Current Study with Previously Reported Ranges in the Literature.

Study Characteristics	Current Study	Literature Range (Caldas et al., 2007; Flores-Mir et al., 2006)	Performance
Inter-Observer (Fleiss κ)	0.611	0.60-0.75	Comparable
Intra-Observer (ICC)	0.973	0.65-0.85	Excellent
Intra-Observer (Cohen's κ)	0.870	0.60-0.80	Superior
Number of Observers	13	2-4	Comprehensive
Institutional Diversity	4 sectors	Single institution	Enhanced
Clinical Protocol	Standardized	Variable/Unreported	Systematic

5.4 Phase 2: Insights from Pilot Modelling and System Development

5.4.1 Limitations of End-to-End Classification for CVM Staging

Pilot experiments using conventional CNN classifiers highlighted systematic limitations in automated CVM staging. Although CNNs can extract global radiographic features, CVM staging depends on subtle morphological changes localised to C2–C4 and, in ambiguous cases, emerging concavity in C5. Because these vertebrae occupy a small portion of the lateral cephalogram, global classifiers must infer skeletal

maturation indirectly from contextual textures and non-diagnostic background structures, which introduces noise and weakens discriminative performance. Transitional stages, particularly CVM II, III and IV, were disproportionately misclassified, mirroring the biological continuity of pubertal acceleration and deceleration phases. This finding is consistent with the broader medical imaging literature, where CNNs struggle when clinically relevant anatomy reflects small, complex, or context-dependent structures embedded in large radiographic fields (Esteva et al., 2021; Litjens et al., 2017). In this context, the primary bottleneck was not the capacity of CNNs per se, but the mismatch between global pixel-level learning and the anatomically localised, ordinal nature of skeletal maturation.

5.4.2 Oversampling as a Diagnostic Probe of Dataset Imbalance

Class imbalance emerged early as a modelling constraint due to limited representation of CVM I and II, reflecting the ethical realities of paediatric radiography, where imaging is performed only when clinically indicated to minimise radiation exposure (European Commission, 2015; McNamara & Franchi, 2018). To explore how imbalance influenced model behaviour, ROS was applied experimentally as a diagnostic probe rather than as a performance optimisation technique. Oversampling improved training stability and reduced volatility in validation performance, consistent with well-documented effects of balancing strategies in supervised learning (Buda et al., 2018; Chawla et al., 2002; He & Garcia, 2009; Johnson & Khoshgoftaar, 2019). However, ROS did not eliminate transitional misclassifications, indicating that imbalance alone could not explain the difficulty observed in mid-pubertal stages. Instead, classifier failure concentrated within the same biological intervals where clinicians experience the greatest staging uncertainty, implying that the computational challenge arises from biological ordinality and anatomical subtlety rather than solely from skewed prevalence. From a clinical standpoint, this alignment reinforces that early growth stages are genuinely rare events in routine practice, rather than artefacts of sampling. From an AI standpoint, the probe clarified that balancing techniques cannot substitute for architectural mechanisms capable of modelling small, ordinal, region-specific structures. Accordingly, oversampling was not retained in the final CVMaXX system to preserve clinically realistic stage frequencies and avoid artificial inflation of performance estimates.

5.4.3 Anatomical Localisation as a Methodological Requirement

The outcomes of the pilot and oversampling experiments indicated that anatomical localisation constitutes a methodological prerequisite for automated CVM staging. In clinical assessment, orthodontists sequentially evaluate vertebral bodies and concavities, focusing on progressive developmental cues rather than global craniofacial context (Flores-Mir et al., 2006; McNamara & Franchi, 2018). Without localisation, CNNs must learn maturation indirectly from global image statistics, making them vulnerable to confounding variation in head posture, soft tissue, and radiographic exposure. Anatomical localisation is increasingly recognised as a necessary component in medical AI systems involving small or context-dependent structures, including retinal microlesions, dental caries, pulmonary nodules and joint spaces (Esteva et al., 2021; Litjens et al., 2017; Schwendicke et al., 2019). In CVM staging, localisation directs the model's representational capacity toward morphologically relevant structures and away from irrelevant contextual textures, thereby aligning computational attention with clinical diagnostic reasoning. This insight motivated the transition from naïve end-to-end classifiers toward a detection-driven pipeline. These methodological considerations establish localisation as a prerequisite for automated CVM staging, which in practice is operationalised through object detection architectures.

5.4.4 Advantages of Object Detection for Transitional Stages

Object detection architectures, operationalised through the YOLOv8 framework, demonstrated substantial advantages over global classifiers. By isolating vertebral regions, detection reduced background interference and amplified morphological signals associated with skeletal maturation, especially during biologically transitional phases. Detection improved sensitivity to ordinal morphological cues, including the deepening of inferior border concavity and proportional elongation of cervical bodies, which play a central role in staging mid-pubertal growth (Flores-Mir et al., 2004; McNamara & Franchi, 2018; Perinetti et al., 2018). From a modelling perspective, detection permits the network to encode local structural relationships, which enhances interpretability and reduces the risk of spurious correlations. These advantages mirror findings from other medical imaging tasks where two-stage detection-classification pipelines outperform global CNNs for anatomically

constrained interpretation (Jocher, 2023; X. Liu et al., 2019; Schwendicke et al., 2019). Collectively, these results support the conclusion that object detection is not merely beneficial but architecturally necessary for clinically plausible CVM staging.

5.4.5 Biological Continuity, Ordinality and Modelling Difficulty

The challenges encountered by both clinicians and automated models in transitional stages stem from a shared constraint: skeletal maturation is continuous, yet CVM staging imposes discrete categorical labels. During the pubertal acceleration phase, vertebral morphology changes rapidly and in a temporally compressed manner, which makes the visual differences between late CVM Stage II and early CVM Stage III or late CVM Stage III and early CVM Stage IV subtle rather than sharply defined (Baccetti et al., 2005; McNamara & Franchi, 2018). When biological transitions occur over short developmental intervals, boundaries become inherently ambiguous. This ambiguity is not an artefact of examiner variability but a genuine expression of continuous growth dynamics (Franchi et al., 2000; Hägg & Taranger, 1982).

Understanding transitional behaviour as an expression of biological continuity has implications for both clinical staging and AI-based CVM systems. When a continuous, ordinal maturation trajectory is forced into discrete categorical labels, boundary artefacts inevitably arise and may be misinterpreted as classification errors. In practice, both clinicians and models tend to struggle at the same transitional boundaries for the same biological reasons, suggesting that variability stems from the nature of the signal rather than from limitations of the observer or the algorithm (Cao et al., 2020; Khazaei et al., 2023; Mathew et al., 2022).

This reframing highlights that modelling difficulty is not simply an engineering challenge but a consequence of discretising a continuous process. Continuous or graded predictions may better capture the ordinal progression of skeletal maturation than rigid categorical labels. Similarly, differentiating early and late transitional phases rather than treating transitions as uniform could align CVM staging more closely with clinical decision-making, where treatment timing depends on whether a patient is approaching, at, or exiting the pubertal growth spurt. These considerations mirror broader discussions in medical imaging, where converting continuous biological processes into discrete

categories introduces uncertainty that cannot be resolved simply by using larger datasets or more complex model architectures (Esteva et al., 2021; Litjens et al., 2017).

In sum, transitional ambiguity should not be interpreted as noise to be eliminated but as biological information that requires careful representation. For orthodontics, acknowledging biological continuity clarifies why transitional stages are clinically important. For computational modelling, it highlights the need for frameworks that can accommodate ordinal structure and biological gradation rather than assuming discrete classification states.

5.4.6 Conceptual Integration for the CVMaXX Framework

Taken together, the pilot experiments indicate that automated CVM staging requires integration of anatomical localisation, developmental ordinality and class-aware learning. CNN classifiers alone are insufficient not due to representational limitations, but because naïve end-to-end architectures assume discrete categories and global morphology. Oversampling clarified that class imbalance exacerbates but does not cause transitional staging difficulty, while detection addressed the need for anatomical grounding and morphological fidelity. These insights provided the architectural rationale for developing the CVMaXX framework and informed subsequent Phase 3 validation. Rather than treating biological ambiguity as a limitation to be eliminated, the modelling process recognised it as a central feature of skeletal maturation and designed system components around that reality.

A notable dimension of these findings concerns their implications for orthodontic workflow. CVM staging informs treatment planning by narrowing the window for growth modification, anchoring the timing of interventions such as functional appliances, suture expansion, or orthopaedic protocols (Angelieri et al., 2015; Franchi et al., 2021; McNamara & Franchi, 2018; Proffit et al., 2018; Tulloch et al., 2004). In routine practice, staging is performed visually at the chairside, but interpretive variability, particularly at transitional stages, can delay or complicate decision-making. By localising anatomical cues and standardising stage outputs, the CVMaXX system has the potential to streamline this component of diagnostic reasoning without displacing clinician judgement. Rather than proposing automation, the system can operate as a supportive triage step within the consultation pathway, providing consistent staging information that is interpretable and clinically actionable.

The overarching value proposition of CVMaXX is not automation but the standardisation of staging in maturational intervals where diagnostic ambiguity influences treatment timing.

5.5 System Reliability and External Validation

Automated skeletal maturation assessment differs from conventional diagnostic classification tasks in that its primary clinical relevance concerns the temporal dimension of treatment planning rather than the identification of a pathological condition. In orthodontics, the utility of maturation staging lies in its capacity to inform when growth-sensitive interventions should be initiated, intensified or deferred, and the consequences of staging are therefore inferential and justificatory rather than purely descriptive (Baccetti et al., 2006; McNamara & Franchi, 2018). External validation and reliability analysis consequently function as translational thresholds that determine whether an automated system participates in the interpretive conditions that govern timing-dependent orthodontic decisions. For CVMaXX, evidence of systematic reliability and structured agreement with expert examiners establishes whether its staging outputs carry the potential to be treated as clinically defensible rather than merely computationally correct, echoing broader expectations for medical AI deployed in decision-support roles in clinical settings (Kelly et al., 2019; Topol, 2019).

5.5.1 Transitional Stages and Interpretive Alignment

External validation showed that CVMaXX exhibited patterned behavioural consistency with trained orthodontic examiners, particularly during maturational intervals that influence treatment timing. Agreement was strongest when vertebral morphology provided clear cues, indicating that model–expert convergence reflected shared interpretive reasoning rather than incidental label matching. This form of convergence is clinically meaningful because staging decisions are rarely made by rigidly applying categorical boundaries; instead, clinicians weigh developmental expectations, growth velocity trajectories, and anticipated treatment windows when determining whether a patient is approaching, within, or past the period of maximal growth modification (Flores-Mir et al., 2006; McNamara & Franchi, 2018; Perinetti et al., 2018). From this perspective, interpretive alignment indicates that CVMaXX

engages in an inferential process with the same justificatory logic clinicians apply when determining whether a patient is sufficiently close to, within or past the peak pubertal growth potential.

Interpreting automated staging in justificatory rather than purely categorical terms aligns with recent discussions in medical AI, which characterise decision-support systems as collaborators in shared interpretive tasks rather than as replacements for expert judgement (Coiera, 2018; Shortliffe & Sepúlveda, 2018). For clinical tasks that depend on developmental timing or cyclical physiological processes, alignment in interpretive behaviour carries more practical weight than perfect categorical agreement, as the priority is whether the inference is defensible within the context of treatment planning. In this sense, the findings suggest that automated staging can participate in orthodontic decision-making not as an oracle that outputs correct labels, but as a tool that produces clinically reasonable inferences about developmental timing. This mirrors how expert examiners integrate skeletal maturity indicators when sequencing interventions such as functional appliances, orthopaedic correction, or surgical planning (O'Brien et al., 2003; Proffit et al., 2018).

Taken together, the pattern of interpretive alignment indicates that CVMaXX internalises morphologic cues in a way that preserves ordinal relations relevant to therapeutic timing. This supports the broader claim that the value of automated CVM staging lies not solely in categorical precision, but in the extent to which its outputs can be treated as reasonable and defensible by clinicians within the practical architecture of growth-sensitive intervention. The implication is that automated staging may enhance the clinician's ability to synthesise maturational indicators rather than merely reproduce categorical classifications, consistent with arguments in the clinical AI literature that emphasise alignment with professional reasoning practices as a prerequisite for successful deployment (Dzau & Ginsburg, 2016).

5.5.2 System Reliability and Agreement with Expert Consensus

Reliability analysis demonstrated that CVMaXX produced reproducible staging outputs under repeated evaluation and that these outputs exhibited substantial agreement with expert consensus according to weighted kappa metrics (Perinetti et al., 2018; Schoretsaniti et al., 2021). Intraclass correlation coefficients further indicated that variance in staging reflected meaningful differences across cases rather than instability

in inference. In the context of translational clinical AI, reliability establishes the epistemic preconditions for downstream deployment, as clinicians require temporally stable behaviour to distinguish physiological developmental change from algorithmic noise (Kelly et al., 2019; Schwendicke et al., 2019).

Strict accuracy quantified exact convergence with expert staging, while inclusive accuracy preserved adjacency relations that clinicians often treat as equivalent for treatment timing decisions. This dual framing mirrors clinical reasoning, in which adjacent maturational stages may be considered functionally interchangeable when they do not alter the therapeutic window for growth modification or functional appliance therapy (Baccetti et al., 2006; Proffit et al., 2018). Notably, this pattern of inclusive agreement suggests that CVMaXX exhibits clinician-like behaviour in its staging outputs, most commonly predicting either the reference stage or an immediately adjacent stage. Such behaviour is consistent with the ordinal and continuous nature of cervical vertebral maturation, where neighbouring stages represent closely related points along a developmental continuum rather than discrete biological states (Baccetti et al., 2005; McNamara & Franchi, 2018). In this sense, uncertainty is expressed within neighbouring stages rather than as gross misclassification, closely reflecting expert reasoning in routine clinical practice.

From a reliability standpoint, ordinal continuity is essential because maturation assessment functions as a temporal instrument, and its justification derives from stability across longitudinal evaluation rather than categorical distinctness at a single time point. Reliability therefore provides the basis upon which automated staging outputs may be integrated into longitudinal monitoring of skeletal maturity, consistent with broader principles of decision-support deployment in clinical environments (Coiera, 2018; Dzau & Ginsburg, 2016).

An additional consideration relates to demographic and population-level equity. Malaysia's multi-ethnic clinical environment introduces inherent variability in craniofacial growth trajectories, skeletal maturation timing, and radiographic presentation. Although the present study was not powered to evaluate subgroup performance, successful external validation against a heterogeneous sample suggests that CVMaXX can accommodate demographic variation encountered in routine clinical settings. Equitable deployment of CVM decision-support tools will ultimately require datasets that reflect the phenotypic diversity of the populations they are intended to serve, underscoring the importance of multi-centre and multi-ethnic contributions to

dataset development in future work. The potential to reduce staging variability across demographic subgroups further situates automated CVM assessment within the broader agenda of fairness and reproducibility in medical AI.

5.5.3 External Validation Across Diverse Clinical Images

External validation provides evidence that CVMaXX can operate under conditions that approximate the realities of orthodontic practice rather than the controlled environment of dataset curation. Clinical radiographs vary in positioning, radiographic exposure, anatomical visibility and soft tissue superimposition, all of which can influence both human and automated staging (Flores-Mir et al., 2006; Schwendicke et al., 2019). The system's stable interpretive behaviour under such heterogeneity indicates that its internal representation of maturation cues generalises beyond idealised imaging conditions (Esteva et al., 2021; Kelly et al., 2019). This form of generalisation is essential for deployment because clinicians routinely interpret cephalograms that differ in quality, clarity and anatomical presentation, and a system that fails to accommodate such variation would have limited applicability.

Demographic factors further complicate staging in real-world settings. Orthodontic patients in Malaysia comprise a multi-ethnic population with Malay, Chinese, Indian and minority ethnic groups contributing to the craniofacial phenotype distribution (Department of Statistics Malaysia, 2023; Khazaei et al., 2023). Craniofacial variation has been shown to influence both skeletal morphology and growth trajectories across populations, yet most CVM staging research has focused on Western cohorts, limiting external validity (Flores-Mir et al., 2004). The observed stability of CVMaXX under multi-ethnic validation suggests that automated staging can accommodate morphological variability that is insufficiently represented in existing datasets. From a patient and clinician perspective, this is consequential because treatment decisions must be defensible across diverse phenotypes encountered in daily practice, not only those represented in research cohorts. From a systems perspective, such generalisability meets expectations that medical AI deployments should contribute to equitable care rather than increasing disparities (Dzau & Ginsburg, 2016; Topol, 2019).

The external validation findings also imply that the system did not overfit idiosyncratic radiographic features from the developmental dataset. Instead, its

behaviour reflects developmental cues that remain interpretable across imaging variations, indicating that the system participates in a clinically meaningful interpretive task rather than reproducing dataset-specific artefacts. In clinical AI, this distinction marks the boundary between a research prototype and a system with potential translational viability (Z. Li & Wang, 2025). For CVM staging in particular, such viability is critical because staging influences decisions whose consequences are borne by patients, families and providers across extended treatment timelines.

5.5.4 Reliability as a Condition for Clinical Deployment

Within orthodontic care, the justification for using skeletal maturation indicators is inherently temporal. Clinicians employ staging to judge whether and when to initiate functional appliances or orthopaedic interventions, monitor the progression of skeletal growth and anticipate whether skeletal correction may require surgical intervention in late adolescence (Baccetti et al., 2005; O'Brien et al., 2003; Proffit et al., 2018). For an automated staging system to participate in such decision making, it must exhibit reliable behaviour across repeated assessments. Reliability ensures that observed changes in staging over time reflect biological development rather than algorithmic variability, a distinction that directly affects clinical interpretation. In external validation, CVMaXX demonstrated reproducibility consistent with this expectation, indicating that its outputs possess the temporal coherence required for monitoring growth trajectories.

From the clinician's perspective, staging constitutes one component of a broader treatment reasoning system that integrates skeletal pattern, dental compensation, airway considerations, growth potential and patient-specific constraints. Automated staging therefore functions as an adjunctive informational input rather than a directive instruction, and its value depends on its ability to support rather than supplant clinical justification (Coiera, 2018; Shortliffe & Sepúlveda, 2018). Reliability is foundational in this regard because inconsistent outputs disrupt treatment planning, increase uncertainty and undermine trust. Conversely, stable outputs allow clinicians to incorporate staging into longitudinal decision making in a manner that preserves professional responsibility and judgement.

From a translational AI standpoint, reliability also represents a regulatory and ethical requirement. Emerging frameworks for medical AI deployment emphasise reproducibility, traceability and interpretability as preconditions for integration into

clinical workflows (Kelly et al., 2019; Topol, 2019). Systems that influence treatment decisions must behave predictably, particularly when those decisions extend across developmental windows and carry consequences for patients and families over months or years. The observed reliability in external validation indicates that CVMaXX meets the preliminary translational benchmarks required for consideration as a decision-support tool in growth-sensitive orthodontic care. While further prospective validation is warranted, these findings support the inference that automated CVM staging can be integrated into treatment planning without compromising the justificatory architecture upon which clinical decision making depends.

5.5.5 Implications for Orthodontic Timing and Treatment Planning

The value of skeletal maturation assessment in orthodontics derives from its capacity to inform the timing of interventions that rely on growth for their therapeutic effect. Functional appliances and orthopaedic treatments demonstrate their greatest skeletal impact during the pubertal acceleration phase, after which dentoalveolar compensation becomes a predominant mode of correction and orthognathic surgery becomes more likely for certain skeletal discrepancies (Baccetti et al., 2005; O'Brien et al., 2003; Proffit et al., 2018). In clinical practice, decisions to initiate or defer growth modification require judgements about whether a patient is approaching, within or past the interval of maximal responsiveness. Automated CVM staging can contribute to this reasoning process if its outputs are considered sufficiently reliable, interpretable and consistent with expert judgement.

From a clinician's standpoint, staging functions as a temporal instrument rather than a categorical diagnosis. It informs not only whether treatment is possible but when it should be initiated, and whether intensification or alternative strategies are indicated. Mistimed interventions may prolong treatment, increase reliance on dental rather than skeletal correction, or, in some cases, shift patients toward surgical management during late adolescence (McNamara & Franchi, 2018; O'Brien et al., 2003; Proffit et al., 2018). From a patient and family perspective, mistiming contributes to treatment burden in the form of increased appliances, prolonged chair time and extended follow-up. Such burdens are meaningful in paediatric care where treatment compliance, schooling schedules, and caregiver availability interact with the logistics of orthodontic appointments (Ghonmode et al., 2022; Santana et al., 2025).

The interpretive alignment and reliability demonstrated by CVMaXX in external validation implies that automated staging could be incorporated as an adjunctive input into timing-sensitive planning decisions. This does not imply replacement of human judgement but reflects the logic of clinical decision-support systems more broadly, in which tools contribute to the informational substrate on which decisions are made rather than determining the decision itself (Coiera, 2018; Shortliffe & Sepúlveda, 2018). When staging information is stable and clinically intelligible, clinicians can employ it to justify initiation, deferment or sequencing decisions, thereby enhancing the defensibility of timing-sensitive interventions. From the provider standpoint, timing affects workflow efficiency, scheduling and treatment outcomes. This may be particularly valuable in transitional maturational intervals where staging uncertainty is greatest and where the consequences of misjudging the growth trajectory are clinically meaningful (Ghonmode et al., 2022; Santana et al., 2025).

5.5.6 Micro- and Macro-Level Oral Health Economic Considerations

Growth modification and timing-sensitive orthodontic care carry economic implications at both micro and macro levels. At the micro level, treatment duration, appliance complexity and the need for retreatment contribute directly to financial cost and indirectly to family burden through caregiver absenteeism, school disruption and logistical coordination (Ghonmode et al., 2022; Santana et al., 2025). Suboptimal treatment timing may reduce the skeletal effects achievable through growth modification and increase reliance on dentoalveolar compensation, particularly in more severe skeletal discrepancies (O'Brien et al., 2003; Proffit et al., 2018). Automated staging systems that increase consistency in maturational assessment therefore hold the potential to mitigate such burdens by supporting more predictable treatment timing and reducing variability in clinical decision making.

At the provider level, orthodontic practices operate within workflow constraints related to chair time, appliance management and scheduling efficiency. Treatment prolongation increases clinical overhead and reduces the capacity to distribute appointments across patient cohorts with differing needs. The introduction of adjunctive decision-support systems that stabilise timing decisions may contribute to more efficient resource allocation, particularly for practitioners managing large paediatric caseloads . Although formal cost analyses were beyond the scope of the present study,

the translational logic aligns with broader movements in dentistry and medicine toward value-based care in which treatment effectiveness, duration and burden are considered in relation to resource expenditure (Coiera, 2018; Dzau & Ginsburg, 2016; FDI World Dental Federation, 2021).

At the macro level, oral health expenditures constitute a measurable component of national health budgets, and orthodontic care intersects with both public and private financing mechanisms. In Malaysia, a mixed public–private system contributes to variability in access to orthodontic services, with timing-sensitive care often situated within private or out-of-pocket models that transfer a significant portion of the economic burden to families (Department of Statistics Malaysia, 2023). For health systems, mistimed intervention can manifest as increased utilisation, deferred surgical care or higher overall treatment burden, all of which contribute to broader oral health expenditure. Automated maturation systems that support more reliable timing decisions could therefore generate downstream efficiencies by reducing retreatment rates, decreasing surgical conversion and enhancing treatment predictability. These hypotheses warrant future evaluation through formal health technology assessment frameworks that quantify economic outcomes, cost-effectiveness and societal burden.

Viewed collectively, the micro- and macro-level considerations suggest that automated CVM staging has potential implications beyond diagnostic accuracy, extending to treatment burden, resource allocation and oral health economics. While preliminary, the present findings indicate that reliability and interpretive alignment establish a basis for considering automated staging within a value-based and efficiency-oriented model of orthodontic care. Such considerations align with contemporary perspectives in translational dental research that emphasise the integration of clinical, economic and systems-level reasoning when evaluating the contribution of innovative technologies to patient care and population health (Coiera, 2018; Dzau & Ginsburg, 2016; FDI World Dental Federation, 2021)

5.6 Phase 3: Methodological Contributions and Performance of CVMaXX

The methodological contribution of this work lies in the design and evaluation of a hybrid architecture that couples precise vertebral localisation with structured feature-based classification. Rather than relying on a single end-to-end convolutional network, CVMaXX integrates YOLOv8 for anatomical localisation of C2–C4 with a

gradient boosting classifier that interprets extracted morphological descriptors. This design reflects the way clinicians first identify the relevant anatomy and then interpret shape changes, and it addresses known limitations of earlier approaches that attempted to infer CVM stage from whole images without explicit anatomical constraints (Litjens et al., 2017; Seo et al., 2021).

Within this framework, LightGBM emerged as the most effective classifier . Its leaf-wise tree growth and gradient-based one-side sampling strategies are well suited to high-dimensional, heterogeneous morphometric data and support efficient modelling of complex non-linear relationships among vertebral features (Ke et al., 2017). In this study, LightGBM achieved near-ceiling discriminative performance, with an AUC approximating 0.98, indicating that higher-stage cases were almost always ranked above lower-stage cases. This performance was consistent across internal testing and external validation, suggesting that the classifier captured generalisable maturation signals rather than dataset-specific artefacts (Mandrekar, 2010; Mathew et al., 2022).

Alternative ensemble methods, particularly XGBoost and Random Forest, provided important comparative baselines (Breiman, 2001; T. Chen & Guestrin, 2016b). XGBoost achieved excellent performance with AUC values above 0.90 and demonstrated strong sensitivity–specificity trade-offs, confirming the suitability of gradient boosting for structured skeletal maturity data (T. Chen & Guestrin, 2016a; Mandrekar, 2010). Random Forest, while somewhat less discriminative, still achieved AUC values consistent with excellent diagnostic performance and served as a robust benchmark for evaluating the added value of more sophisticated boosting methods (Breiman, 2001; Litjens et al., 2017). The consistent superiority of LightGBM across AUC and related performance indicators, supported by DeLong's pairwise comparisons, justified its selection as the primary classifier for CVMaXX (DeLong et al., 1988).

Performance metrics were interpreted with attention to both statistical and clinical meaning. Overall accuracy was high but insufficient as a stand-alone indicator due to stage imbalance, particularly the relative scarcity of early CVM stages, which reflects ethical constraints on radiographic exposure in younger children (Khazaei et al., 2023; Sokolova & Lapalme, 2009). Stage-wise precision and recall revealed that the system was most confident and sensitive in morphologically distinct stages, whereas transitional stages displayed lower values, mirroring human diagnostic difficulties and

reinforcing the biological continuity of pubertal growth (Nestman et al., 2011; Perinetti et al., 2012; Powers, 2020).

F1-scores and confusion matrix analysis provided a more nuanced view of error patterns. Misclassifications were largely confined to adjacent stages rather than non-adjacent or clinically implausible errors, and macro- and weighted F1-scores were comparable to or exceeded those reported in previous AI-based CVM studies (Mathew et al., 2022; Seo et al., 2021). This adjacency-bounded error pattern indicates that residual inaccuracies arise primarily in biologically ambiguous intervals and supports the interpretation that the classifier respects the ordinal structure of skeletal maturation.

Agreement metrics further contextualised performance relative to expert consensus. Unweighted Cohen's κ provided a conservative estimate of case-level agreement, appropriate for a primary validation objective focused on whether the system could match expert staging at the overall case level (McHugh, 2012; Viera & Garrett, 2005). Supplementary analyses with linear and quadratic weighted κ revealed even higher concordance, consistent with the predominance of minor, adjacent-stage disagreements. Together with intraclass correlation coefficients, these metrics indicate that CVMaXX behaves comparably to calibrated human examiners and that disagreements are biologically bounded rather than arbitrary (Cohen, 1968; Fleiss et al., 1969; Nestman et al., 2011; Perinetti et al., 2012).

Viewed in combination, these findings show that the hybrid YOLOv8–LightGBM architecture balances high discriminative performance with clinically interpretable behaviour. The system achieves expert-comparable staging accuracy, concentrates errors in biologically ambiguous transitional intervals, and provides reproducible outputs suitable for downstream clinical and translational interpretation.

5.7 Clinical Relevance, Training Value, and Digital Health Context

The clinical relevance of CVMaXX lies in its ability to provide reproducible, expert-comparable staging that directly informs timing-sensitive orthodontic decisions. In morphologically distinct stages, where concavity and vertebral shape are clearly differentiated, the system reached high levels of exact agreement with expert examiners. This reduces inter-observer variability and supports more standardised staging across clinicians and institutions (Baccetti et al., 2005; Perinetti et al., 2012). For clinicians,

such standardisation can enhance confidence in growth modification timing and facilitate communication when co-managing cases between centres.

Transitional stages, particularly CVM Stages II and IV, remained challenging, with lower accuracy and modest increases in adjacent-stage misclassification. However, this pattern is consistent with the well-documented difficulties that human examiners face in these intervals and reflects intrinsic biological ambiguity rather than purely algorithmic failure (Nestman et al., 2011; Perinetti et al., 2012). From a practical standpoint, this reinforces the importance of interpreting transitional-stage predictions cautiously and integrating them with other clinical indicators, such as skeletal pattern, growth history and treatment objectives. In these cases, CVMaXX is best conceptualised as an adjunct that highlights likely maturational status and supports, rather than replaces, clinician judgement.

Beyond direct casework, CVMaXX offers potential value in training and standardisation. Trainees often struggle to recognise subtle morphological changes and to apply CVM criteria consistently, leading to substantial variability in early assessments (Nestman et al., 2011). By providing a consistent reference and visually interpretable classifications, CVMaXX can function as a teaching aid that helps learners align their mental models of vertebral morphology with consensus-based standards. Similar applications of AI as calibration tools have been reported in radiology and other imaging domains, where automated systems are used to benchmark trainee interpretations and support skill development (Esteva et al., 2021).

From a digital health perspective, the modular design of CVMaXX supports deployment in diverse clinical environments, including settings with limited specialist access. A web-based interface would allow clinicians to upload lateral cephalograms and receive standardised staging outputs, extending specialist-level assessment to rural or underserved regions without requiring local expertise (Flores-Mir et al., 2006; Topol, 2019; Wiens et al., 2019). In high-volume clinics, the same infrastructure could reduce examiner fatigue and optimise workflow by providing rapid second opinions on maturation status. These potential applications align with emerging digital health frameworks that emphasise AI as augmentative, interpretable and embedded within existing clinical reasoning pathways rather than as stand-alone diagnostic tools (Coiera, 2018; Ratitch et al., 2023; Shortliffe & Sepúlveda, 2018).

5.8 Limitations of the Present Study

Several limitations should be recognised when interpreting the findings of this study. First, early CVM stages, particularly CVM Stages I and II, were under-represented in the dataset. This reflects ethical and clinical constraints, as radiographic imaging of younger children cannot be justified solely for research and must be based on clinical need (European Commission, 2015; Khazaei et al., 2023). Consequently, model estimates in early stages are informed by fewer examples and may exhibit greater uncertainty. Misclassifications were most frequent at transitional boundaries, particularly between CVM Stages II–III and IV–V, where progressive morphological change produces overlapping features (McNamara & Franchi, 2018). Reduced agreement in these intervals mirrors prior studies and reflects intrinsic biological variability rather than purely technical shortcomings (Baccetti et al., 2005; Nestman et al., 2011; Perinetti et al., 2012).

Second, the study relied exclusively on two-dimensional lateral cephalograms. While cephalograms remain the clinical standard because of their accessibility and comparatively low radiation dose, they lack three-dimensional anatomical information, and subtle variations in vertebral depth or spatial orientation may be obscured, especially in borderline cases (Shim et al., 2012). Emerging evidence suggests that three-dimensional modalities, such as CBCT, can improve morphological characterisation, although routine use remains constrained by radiation exposure and cost (Echevarría-Sánchez et al., 2020b).

Third, the dataset was retrospective and derived from a limited number of institutions within a single geographic region. Although this provided consistency in imaging and annotation protocols, it also restricts generalisability. Craniofacial morphology and growth patterns vary across populations, and models trained on relatively homogeneous datasets may encode population-specific biases (Flores-Mir et al., 2006; Khazaei et al., 2023; Topol, 2019; Wiens et al., 2019). External validation across multi-centre, multi-ethnic cohorts is therefore necessary before widespread deployment. Although subgroup analyses were not performed, future work should examine demographic variation in model performance given Malaysia's multi-ethnic clinical context.

Fourth, although data balancing strategies were used during pilot development to mitigate class imbalance, oversampling can increase the risk of overfitting by

replicating minority class samples without introducing new anatomical variation (Mathew et al., 2022). While later phases of the study emphasised balanced, curated datasets without reliance on synthetic oversampling, residual risks of subtle overfitting cannot be fully excluded.

Finally, the present work is cross-sectional and focuses on stage classification at single time points. Orthodontic treatment planning, however, often relies on predicting future growth trajectories rather than static stage identification (Adarsh et al., 2025; Baccetti et al., 2005; Franchi et al., 2013). Longitudinal predictive validity was not assessed and remains an important limitation for fully growth-based decision support.

5.9 Future Directions and Clinical Translation

The translational implications of the present findings extend beyond algorithmic performance. By integrating calibrated annotation, anatomical localisation, ordinal classification, and external validation, the CVMaXX system addresses several recognised barriers in medical AI, including biological interpretability, dataset heterogeneity, workflow alignment, and equity considerations. In orthodontics, where diagnostic uncertainty at transitional maturational stages influences treatment timing and modality selection, the capacity to standardise staging decisions may improve both clinical consistency and patient counselling. Within the broader digital health context, the study exemplifies how AI-enabled decision support can augment specialised diagnostic tasks without displacing clinician expertise. Further translational work will be required to assess usability, generalisability, and cost-benefit considerations in real-world environments, but the findings indicate a credible trajectory toward clinically meaningful deployment.

Future work should extend CVMaXX along several complementary trajectories. First, a key research priority concerns the handling of transitional morphology. Future studies may benefit from differentiating early and late transitional phases rather than treating the CVM Stage II–III and CVM Stage III–IV boundaries as uniform. A graded or continuous representation of transitional morphology would better reflect the underlying biology of the pubertal growth spurt and may enhance both treatment-timing decisions and model performance. Operationalising such distinctions would require

expanded datasets with sufficient representation across transitional intervals and may encourage probabilistic or ordinal frameworks in place of strict nominal labels.

Contemporary orthodontic literature continues to support the clinical relevance of skeletal maturation assessment in treatment timing, particularly for growth-modifying interventions in growing patients. Recent systematic reviews and evidence syntheses suggest that treatment initiated during periods of active skeletal growth may enhance the skeletal contribution of functional and orthopaedic therapy, although treatment outcomes remain multifactorial and influenced by patient compliance, treatment modality, and individual growth variability (Perinetti & Contardo, 2017). In Class II malocclusion management, skeletal maturity assessment remains commonly incorporated into treatment planning to optimise the timing of functional appliance therapy and maximise growth responsiveness where possible (Franchi et al., 2021; Khade et al., 2023). Similarly, early orthopaedic intervention in selected Class III cases may improve maxillary protraction effects and reduce the severity of skeletal discrepancy during growth (Ngan & Moon, 2015). Nevertheless, recent evidence also emphasises that skeletal maturation indicators should be interpreted cautiously and as adjunctive tools within a broader clinical assessment framework rather than as absolute predictors of treatment success (Pacha et al., 2023). Longitudinal validation is also needed to determine whether automated staging can reliably predict growth patterns and support timing decisions for Class II and Class III treatment. Serial imaging studies correlating CVMaXX-derived stages with subsequent skeletal and dental outcomes would clarify its utility for personalised treatment timing (Adarsh et al., 2025; Franchi et al., 2013).

Second, multimodal integration offers an important opportunity to enhance predictive robustness. Combining cephalometric CVM staging with dental development indices, anthropometric growth markers, chronological age and selectively acquired three-dimensional imaging could more closely reflect the multi-indicator reasoning employed in clinical practice (Echevarría-Sánchez et al., 2020b; Flores-Mir et al., 2006; Perinetti et al., 2012). Such integration would better capture the complexity of skeletal maturation and improve resilience in transitional stages.

Third, broader external validation across geographically and ethnically diverse populations is essential for fairness and generalisability. Collaborative multi-centre studies would enable assessment of performance across different craniofacial patterns,

imaging protocols and health system contexts, consistent with current recommendations for medical AI deployment (Schwendicke et al., 2019; Topol, 2019; Wiens et al., 2019)

Finally, future iterations of CVMaXX should emphasise hybrid human–AI interfaces and explainability. Visualisation tools such as saliency maps, feature importance plots and region-based heatmaps could help clinicians verify that predictions are driven by anatomically meaningful regions, reinforcing trust and facilitating adoption (Selvaraju et al., 2017; Tonekaboni et al., 2019; Topol, 2019). Interactive interfaces that position clinicians as final decision makers, with the ability to adjust thresholds and review underlying evidence, align with best practices for responsible AI deployment in clinical environments.

5.10 Integrated Synthesis of Findings

Collectively, the findings of this study show that the longstanding challenges of subjectivity and variability in CVM staging can be mitigated through structured calibration and carefully designed AI architectures. The multi-expert calibration framework demonstrated that CVM staging can achieve substantial inter-observer agreement and excellent intra-observer reproducibility when operational definitions are standardised, particularly at transitional boundaries (Nestman et al., 2011; Perinetti et al., 2012). This establishes a reliable annotation foundation upon which automated systems can be trained, addressing a recognised bottleneck in medical AI, where label quality and consistency frequently constrain performance, robustness, and external validity as much as algorithm choice (Schwendicke et al., 2019)

On this foundation, the CVMaXX framework demonstrated that decoupling anatomical localisation and morphological classification yields both methodological and clinical advantages. Anatomical localisation using YOLOv8 restricts analysis to biologically relevant structures, while LightGBM-based classification interprets morphology in a manner that aligns with expert reasoning (Jocher, 2023; Ke et al., 2017). The system achieved expert-comparable strict accuracy, high inclusive accuracy on external validation, and residual misclassifications largely confined to biologically ambiguous transitional stages, consistent with the intrinsic overlap described in developmental morphology literature (Baccetti et al., 2005; Nestman et al., 2011; Perinetti et al., 2012). These outcomes indicate that remaining classification challenges

reflect intrinsic limits of morphology-based staging rather than fundamental algorithmic deficits.

From a clinical perspective, the findings indicate strong translational potential. Positioning the cervical vertebrae as the explicit computational foreground is therefore not a modelling preference but a clinically aligned necessity. By reducing background anatomical noise and constraining the learning task to discriminative vertebral morphology, the localisation-first pipeline mirrors the perceptual strategy clinicians employ when staging skeletal maturation (Baccetti et al., 2005).

Importantly, the architectural shift was not a data volume problem to be addressed by deeper networks or larger datasets, but a task geometry problem in which naïve end-to-end models trained on full cephalograms were required to infer fine-grained vertebral morphology from radiographs dominated by diagnostically irrelevant structures. Localisation reduced this anatomical noise floor and allowed downstream classification to operate on signal rather than confounders, an alignment highlighted in broader medical imaging research linking anatomical grounding to improved robustness, interpretability, and clinical reliability (Esteva et al., 2021; Litjens et al., 2017; Rajpurkar et al., 2018).

Viewed through a clinical-practical lens, this alignment enhances workflow relevance. Orthodontic decision-making around treatment timing and appliance selection is concentrated in transitional maturational intervals where diagnostic ambiguity is greatest (Proffit et al., 2018). A system that standardises staging within these windows offers tangible utility in busy or resource-constrained environments, including private practices and mixed-experience clinical settings, where access to specialist judgement may be limited and throughput pressures elevate the value of reproducibility (Flores-Mir et al., 2004; Topol, 2019).

CHAPTER 6
CONCLUSION

6.1 Conclusion

First, the study addressed the question of whether CVM staging can achieve acceptable reliability among orthodontic experts. The findings demonstrated that, following structured calibration and consensus-based decision protocols, CVM staging achieved substantial inter-observer agreement and excellent intra-observer reproducibility. This confirms that examiner variability, particularly in transitional stages, can be significantly reduced through systematic methodological control (Table 6.1).

Second, the study examined whether a reliable ground truth dataset could be established for artificial intelligence development. The multi-expert calibration framework successfully produced a validated and reproducible reference standard, satisfying a critical prerequisite for supervised machine learning in medical imaging.

Third, the research evaluated whether a fully automated system could accurately localise and classify CVM stages. The proposed CVMaXX system successfully achieved this objective through a two-stage pipeline, with YOLOv8 providing precise localisation of C2 to C4 vertebrae and LightGBM delivering robust morphology-based classification. The proposed system demonstrated high inclusive accuracy during external validation, while maintaining conservative performance under strict evaluation criteria.

Finally, the study established the clinical relevance and validity of the automated framework. The results confirmed that CVMaXX provides reproducible, standardised CVM staging with expert-level performance, supporting its role as a clinically meaningful decision-support tool rather than a replacement for clinician judgment.

In conclusion, this research demonstrates that automated CVM classification can achieve expert-comparable performance when supported by rigorous reliability frameworks and thoughtfully designed artificial intelligence architectures. The CVMaXX system represents a clinically informed, methodologically robust, and scalable solution for skeletal maturity assessment. By directly addressing all research objectives and research questions, this thesis establishes a strong foundation for future

work in multimodal integration, external validation, and longitudinal growth prediction, positioning CVMaXX as a methodologically robust and clinically credible advancement in AI-assisted orthodontic diagnostics.

Table 6.1
Mapping of Research Objectives and Research Questions to Tables and Figures

RO / RQ	Key Conclusion	Quantification	Visual	Clinical Interpretation
RO1 / RQ1	Substantial inter-observer agreement and excellent intra-observer reproducibility when structured calibration and decision protocols	ICC, κ	Figures 4.2–4.3 Table 3.8, 4.8 Section 4.2	Section 5.3
RO2 / RQ2	Reliable expert-derived ground truth can be established through multi-expert calibration, consensus rules, and operational examiner selection.	Agreement metrics	Figure 3.1 Table 3.4-3.5, Figure 4.4, Section 4.2.3	Section 5.3
RO3 / RQ3	Successfully performs automated vertebral localisation and morphology-driven CVM classification with high accuracy.	Accuracy, AUC	Figures 3.6, 4.14–4.16, Table 4.10 Section 4.3-4.5	Section 5.4–5.6
RO4 / RQ4	Expert-comparable diagnostic performance with preserved generalisability on independent unseen datasets.	External validation	Figures 4.7, 4.19 Table 4.11 Section 4.6	Section 5.4, 5.6

REFERENCES

Abdulla Alkhal, H., K Wong, R. W., & Bakr Rabie, A. M. (2008). Correlation between Chronological Age, Cervical Vertebral Maturation and Fishman's Skeletal Maturity Indicators in Southern Chinese. *Angle Orthodontist*, *78*(4), 591–596. https://doi.org/10.2319/051607-235.1

Abdulqader, A. A., Jiang, F., Almaqrami, B. S., Cheng, F., Yu, J., Qiu, Y., & Li, J. (2025). Deep learning approaches for quantitative and qualitative assessment of cervical vertebral maturation staging systems. *PLOS One*, *20*(5), e0323776. https://doi.org/10.1371/journal.pone.0323776

Adarsh, K., Goyal, J. D., Kaur, G. P., Neeharika, T., Samal, S., Datla, P. K. V., & Ravuri, P. (2025). Longitudinal Study on the Impact of Orthodontic Treatment Timing on Skeletal and Dental Development in Adolescents. *Journal of Pharmacy and Bioallied Sciences*, *17*(Suppl 1), S460–S462. https://doi.org/10.4103/jpbs.jpbs_1425_24

Adil, S., Ul Islam, Z., Karim, K., Ali Raza, H., Tauheed, S., & Hassan, S. (2022). Correlation of Dental and Skeletal Maturity Using Dimirgian and Cervical Vertebral Maturation Indices in Orthodontic Patients. *Journal of Gandhara Medical and Dental Science*, *9*(4), 20–24. https://doi.org/10.37762/jgmds.9-4.312

Agarwal, S., & Agarwal, S. (2022). Bone Age Assessment from Lateral Cephalograms Using Deep Learning Algorithms in the Indian Population. *Indian Journal of Dental Research*, *33*(4), 402–407. https://doi.org/10.4103/ijdr.ijdr_955_21

Akay, G., Akcayol, M. A., Özdem, K., & Güngör, K. (2023). Deep convolutional neural network—the evaluation of cervical vertebrae maturation. *Oral Radiology*, *39*(4), 629–638. https://doi.org/10.1007/s11282-023-00678-7

Alhamady, A. M., Ishaq, R. A. R., Alhammadi, M. S., Almashraqi, A. A., & Alhashimi, N. (2024). Evaluation of an objective staging system for assessment of cervical vertebral maturation. *BMC Oral Health*, *24*(1), 97. https://doi.org/10.1186/s12903-023-03844-9

Ali, M. M., Muhammad, N., Muhammad, A., Gulsana, H., Muhammad Hassaan, S., & Mazhar, H. (2024). Interobserver Reliability of Cervical Vertebral Maturation Staging. *Journal of Saidu Medical College, Swat*, *14*(1), 41–44. https://doi.org/10.52206/jsmc.2024.14.1.820

Amasya, H., Cesur, E., Yıldırım, D., & Orhan, K. (2020). Validation of cervical vertebral maturation stages: Artificial intelligence vs human observer visual analysis. *American Journal of Orthodontics and Dentofacial Orthopedics*, *158*(6), e173–e179. https://doi.org/10.1016/j.ajodo.2020.08.014

Amasya, H., Yildirim, D., Aydogan, T., Kemaloglu, N., & Orhan, K. (2020). Cervical vertebral maturation assessment on lateral cephalometric radiographs using artificial intelligence: Comparison of machine learning classifier models. *Dentomaxillofacial Radiology*, *49*(5), 1–28. https://doi.org/10.1259/dmfr.20190441

Angelieri, F., Franchi, L., Cevidanes, L. H. S., & McNamara, J. A. (2015). Diagnostic performance of skeletal maturity for the assessment of midpalatal suture maturation. *American Journal of Orthodontics and Dentofacial Orthopedics*, *148*(6), 1010–1016. https://doi.org/10.1016/j.ajodo.2015.06.016

Atici, S. F., Ansari, R., Allareddy, V., Suhaym, O., Cetin, A. E., & Elnagar, M. H. (2022). Fully automated determination of the cervical vertebrae maturation stages using deep learning with directional filters. *PLoS ONE*, *17*(7), 1–19. https://doi.org/10.1371/journal.pone.0269198

Atici, S. F., Ansari, R., Allareddy, V., Suhaym, O., Cetin, A. E., & Elnagar, M. H. (2023). AggregateNet : A deep learning model for automated classification of cervical vertebrae maturation stages . *Orthodontics & Craniofacial Research*. https://doi.org/10.1111/ocr.12644

Baccetti, T., Franchi, L., De Toffol, L., Ghiozzi, B., & Cozza, P. (2006). The diagnostic performance of chronologic age in the assessment of skeletal maturity. *Progress in Orthodontics*, *7*(2), 176–188.

Baccetti, T., Franchi, L., & Mcnamara, J. A. (2002). An Improved Version of the Cervical Vertebral Maturation (CVM) Method for the Assessment of Mandibular Growth. *Angle Orthodontist*, *72*(4), 316–323.

Baccetti, T., Franchi, L., & McNamara, J. A. (2005). The Cervical Vertebral Maturation (CVM) method for the assessment of optimal treatment timing in dentofacial orthopedics. *Seminars in Orthodontics*, *11*(3), 119–129. https://doi.org/10.1053/j.sodo.2005.04.005

BK, R., P, S., Patil, P., Krishnamurthy, A., DR, M., & S, D. (2025). Development and validation of an artificial intelligence algorithm for cervical vertebral maturation

staging using lateral cephalograms. *Medical Journal Armed Forces India*, *81*(6), 672–679. https://doi.org/10.1016/j.mjafi.2025.08.012

Bowden, B. D. (1976). Epiphysial changes in the hand/wrist area as indicators of adolescent stage. *Australian Orthodontic Journal*, *4*(3), 87–104.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brickley, M. R., Shepherd, J. P., & Armstrong, R. A. (1998). Neural networks: a new technique for development of decision support systems in dentistry. In *Journal of Dent&y* (Vol. 26, Issue 4).

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Byun, B. R., Kim, Y. Il, Yamaguchi, T., Maki, K., Ko, C. C., Hwang, D. S., Park, S. B., & Son, W. S. (2015). Quantitative skeletal maturation estimation using cone-beam computed tomography-generated cervical vertebral images: a pilot study in 5- to 18-year-old Japanese children. *Clinical Oral Investigations*, *19*(8), 2133–2140. https://doi.org/10.1007/s00784-015-1415-6

Caldas, M. de P., Ambrosano, G. M. B., & Haiter Neto, F. (2010). Computer-assisted analysis of cervical vertebral bone age using cephalometric radiographs in Brazilian subjects. *Brazilian Oral Research*, *24*(1), 120–126. https://doi.org/10.1590/S1806-83242010000100020

Caldas, M. de P., Ambrosano, G. M. B., & Haiter-Neto, F. (2007). Use of cervical vertebral dimensions for assessment of children growth. *Journal of Applied Oral Science : Revista FOB*, *15*(2), 144–147. https://doi.org/10.1590/s1678-77572007000200014

Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, *140*, 325–331. https://doi.org/10.1016/j.patrec.2020.11.008

Cavallo, F., Mohn, A., Chiarelli, F., & Giannini, C. (2021). Evaluation of Bone Age in Children: A Mini-Review. In *Frontiers in Pediatrics* (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fped.2021.580314

Cericato, G. O., Bittencourt, M. A. V., & Paranhos, L. R. (2015). Validity of the assessment method of skeletal maturation by cervical vertebrae: A systematic

review and meta-analysis. *Dentomaxillofacial Radiology*, *44*(4). https://doi.org/10.1259/dmfr.20140270

Chatzigianni, A., & Halazonetis, D. J. (2009). Geometric morphometric evaluation of cervical vertebrae shape and its relationship to skeletal maturation. *American Journal of Orthodontics and Dentofacial Orthopedics*, *136*(4), 481.e1-481.e9. https://doi.org/10.1016/j.ajodo.2009.04.017

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, L.-L., Xu, T.-M., Jiang, J.-H., Zhang, X.-Z., & Lin, J.-X. (2008). Quantitative cervical vertebral maturation assessment in adolescents with normal occlusion: A mixed longitudinal study. *American Journal of Orthodontics and Dentofacial Orthopedics*, *134*(6), 720.e1-720.e7. https://doi.org/10.1016/j.ajodo.2008.03.014

Chen, S., Wang, L., Li, G., Wu, T. H., Diachina, S., Tejera, B., Kwon, J. J., Lin, F. C., Lee, Y. T., Xu, T., Shen, D., & Ko, C. C. (2020). Machine learning in orthodontics: Introducing a 3D auto-segmentation and auto-landmark finder of CBCT images to assess maxillary constriction in unilateral impacted canine patients. *Angle Orthodontist*, *90*(1), 77–84. https://doi.org/10.2319/012919-59.1

Chen, T., & Guestrin, C. (2016a). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., & Guestrin, C. (2016b). XGBoost: A scalable tree boosting system. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y. W., Stanley, K., & Att, W. (2020). Artificial intelligence in dentistry: current applications and future perspectives. *Quintessence International (Berlin, Germany : 1985)*, *51*(3), 248–257. https://doi.org/10.3290/j.qi.a43952

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

Coiera, E. (2018). The fate of medicine in the time of AI. *The Lancet*, *392*(10162), 2331–2332. https://doi.org/10.1016/S0140-6736(18)31925-1

Cummaudo, M., De Angelis, D., Magli, F., Minà, G., Merelli, V., & Cattaneo, C. (2021). Age estimation in the living: A scoping review of population data for skeletal and dental methods. In *Forensic Science International* (Vol. 320). Elsevier Ireland Ltd. https://doi.org/10.1016/j.forsciint.2021.110689

Cunha, A. C., Cevidanes, L. H. S., Sant'Anna, E. F., Guedes, F. R., Luiz, R. R., McNamara, J. A., Franchi, L., & Ruellas, A. C. O. (2018). Staging hand-wrist and cervical vertebrae images: A comparison of reproducibility. *Dentomaxillofacial Radiology*, *47*(5). https://doi.org/10.1259/dmfr.20170301

Dadgar, S., Hadian, H., Ghobadi, M., Sobouti, F., & Rakhshan, V. (2021). Correlations among chronological age, cervical vertebral maturation index, and Demirjian developmental stage of the maxillary and mandibular canines and second molars. *Surgical and Radiologic Anatomy*, *43*(1), 131–143. https://doi.org/10.1007/s00276-020-02541-4

De Sanctis, V., Soliman, A., Fiscina, B., Elsedfy, H., Elalaily, R., Yassin, M., Skordis, N., Di Maio, S., Piacentini, G., & Kholy, M. (2014). Endocrine check-up in adolescents and indications for referral: A guide for health care providers. *Indian Journal of Endocrinology and Metabolism*, *18*(7), 26. https://doi.org/10.4103/2230-8210.145055

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837–845.

Demirjian, A., Buschang, P. H., Tanguay, R., & Patterson, D. K. (1985). Interrelationships among measures of somatic, skeletal, dental, and sexual maturity. *American Journal of Orthodontics*, *88*(5), 433–438. https://doi.org/10.1016/0002-9416(85)90070-3

Demirjian A, G. H. T. JM. (1973). A New System of Dental Age Assessment. *Human Biology*, *2*(45), 211–227.

Department of Statistics Malaysia. (2023). *Current Population Estimates, Malaysia, 2023*. http://www.dosm.gov.my

Dhiman, S., Maheshwari, S., & Verma, S. K. (2015). Assessment of maturity in orthodontics: A review. *Journal of Advanced Clinical & Research Insights*, *2*, 100–103. https://doi.org/10.15713/ins.jcri.54

Dipalma, G., Inchingolo, A. D., Inchingolo, A. M., Piras, F., Carpentiere, V., Garofoli, G., Azzollini, D., Campanelli, M., Paduanelli, G., Palermo, A., & Inchingolo, F.

(2023). Artificial Intelligence and Its Clinical Applications in Orthodontics: A Systematic Review. *Diagnostics*, *13*(24), 1–22. https://doi.org/10.3390/diagnostics13243677

Du, W., Bi, W., Liu, Y., Zhu, Z., Tai, Y., & Luo, E. (2024). Machine learning-based decision support system for orthognathic diagnosis and treatment planning. *BMC Oral Health*, *24*(1), 286. https://doi.org/10.1186/s12903-024-04063-6

Dzau, V. J., & Ginsburg, G. S. (2016). Realizing the Full Potential of Precision Medicine in Health and Health Care. *JAMA*, *316*(16), 1659. https://doi.org/10.1001/jama.2016.14117

Echevarría-Sánchez, G., Arriola-Guillén, L. E., Malpartida-Carrillo, V., Tinedo-López, P. L., Palti-Menendez, R., & Guerrero, M. E. (2020a). Reliability of cephalograms derived of cone beam computed tomography versus lateral cephalograms to estimate cervical vertebrae maturity in a Peruvian population: A retrospective study. *International Orthodontics*, *18*(2), 258–265. https://doi.org/10.1016/j.ortho.2020.01.001

Echevarría-Sánchez, G., Arriola-Guillén, L. E., Malpartida-Carrillo, V., Tinedo-López, P. L., Palti-Menendez, R., & Guerrero, M. E. (2020b). Reliability of cephalograms derived of cone beam computed tomography versus lateral cephalograms to estimate cervical vertebrae maturity in a Peruvian population: A retrospective study. *International Orthodontics*, *18*(2), 258–265. https://doi.org/10.1016/j.ortho.2020.01.001

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., & Socher, R. (2021). Deep learning-enabled medical computer vision. *Npj Digital Medicine*, *4*(1), 5. https://doi.org/10.1038/s41746-020-00376-2

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

European Commission. (2015). *Radiation Protection No. 172: Guidelines on Medical Imaging for Children*.

FDI World Dental Federation. (2021). *Vision 2030: Delivering Optimal Oral Health for All*. https://www.fdiworlddental.org/vision2030

Felemban, N. H. (2017). Korelacija izmedu stadija zrelosti cervikalne kralježnice i dentalne dobi u saudijskom uzorku. *Acta Stomatologica Croatica*, *51*(4), 283–289. https://doi.org/10.15644/asc51/4/2

Ferrillo, M., Curci, C., Roccuzzo, A., Migliario, M., Invernizzi, M., & De Sire, A. (2021). Reliability of cervical vertebral maturation compared to hand-wrist for skeletal maturation assessment in growing subjects: A systematic review. In *Journal of Back and Musculoskeletal Rehabilitation* (Vol. 34, Issue 6, pp. 925–936). IOS Press BV. https://doi.org/10.3233/BMR-210003

Firouzinia, M., Khafri, S., Mirzaie, M., Abesi, F., & Hamzeh, M. (2022). Correlation between Skeletal Maturation and Developmental Stages of Canines and Second Molars among Iranian Population. *J Dent Shiraz Univ Med Sci*, *23*(2), 95–101. https://doi.org/10.30476/DENTJODS.2021.87505.1266

Fishman, LS. (1982). Radiographic evaluation of skeletal maturation. A clinically oriented method based on hand-wrist films. . *Angle Orthodontics*, *52*, 88–112.

Fleiss, J. L. (1971). MEASURING NOMINAL SCALE AGREEMENT AMONG MANY RATERS 1. In *Psychological Bulletin* (Vol. 76, Issue 5).

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*(5), 323–327. https://doi.org/10.1037/h0028106

Flieger, R., Matys, J., & Dominiak, M. (2018). The best time for orthodontic treatment for Polish children based on skeletal age analysis in accordance to refund policy of the Polish National Health Fund (NFZ). *Advances in Clinical and Experimental Medicine*, *27*(10), 1377–1382. https://doi.org/10.17219/acem/69976

Flores-Mir, C., Corr, ;, Burgess, A., Champney, M., Jensen, R. J., Pitcher, M. R., & Major, P. W. (2006). Correlation of Skeletal Maturation Stages Determined by Cervical Vertebrae and Hand-wrist Evaluations. In *Angle Orthodontist* (Vol. 76, Issue 1).

Flores-Mir, C., Orth, C., Nebbe, B., & Major, P. W. (2004). Use of Skeletal Maturation Based on Hand-Wrist Radiographic Analysis as a Predictor of Facial Growth: A Systematic Review. In *Angle Orthodontist* (Vol. 74, Issue 1).

Franchi, L., Baccetti, T., & McNamara, J. A. (1999). Treatment and posttreatment effects of acrylic splint Herbst appliance therapy. *American Journal of Orthodontics and Dentofacial Orthopedics*, *115*(4), 429–438. https://doi.org/10.1016/S0889-5406(99)70264-7

Franchi, L., Baccetti, T., & McNamara, J. A. (2000). Mandibular growth as related to cervical vertebral maturation and body height. *American Journal of*

Orthodontics and Dentofacial Orthopedics, *118*(3), 335–340. https://doi.org/10.1067/mod.2000.107009

Franchi, L., Nieri, M., McNamara, J. A., & Giuntini, V. (2021). Predicting mandibular growth based on CVM stage and gender and with chronological age as a curvilinear variable. *Orthodontics and Craniofacial Research*, *24*(3), 414–420. https://doi.org/10.1111/ocr.12457

Franchi, L., Pavoni, C., Faltin, K., McNamara, J. A., & Cozza, P. (2013). Long-term skeletal and dental effects and treatment timing for functional appliances in Class II malocclusion. *The Angle Orthodontist*, *83*(2), 334–340. https://doi.org/10.2319/052912-450.1

Gabriel, D. B., Southard, K. A., Qian, F., Marshall, S. D., Franciscus, R. G., & Southard, T. E. (2009). Cervical vertebrae maturation method: Poor reproducibility. *American Journal of Orthodontics and Dentofacial Orthopedics*, *136*(4), 478.e1-478.e7. https://doi.org/10.1016/j.ajodo.2007.08.028

García-Pola, M., Pons-Fuster, E., Suárez-Fernández, C., Seoane-Romero, J., Romero-Méndez, A., & López-Jornet, P. (2021). Role of Artificial Intelligence in the Early Diagnosis of Oral Cancer. A Scoping Review. *Cancers*, *13*(18), 4600. https://doi.org/10.3390/cancers13184600

Ghonmode, S., Shrivastava, S., Kadaskar, A. R., & Bapat, S. (2022). Socioeconomic burden of orthodontic treatment: a systematic review. *Medicine and Pharmacy Reports*, *96*(2), 154–163. https://doi.org/10.15386/mpr-2457

Gowri Priya, S., & Thirugnanam, A. (2024). Automation in orthodontics and orthopedics using artificial intelligence. In *Advances in Artificial Intelligence* (pp. 359–381). Elsevier. https://doi.org/10.1016/B978-0-443-19073-5.00011-2

Gudhimella, S., Gandhi, V., Glatt, B., Wong, A., Wood, C., Janakiraman, N., Spiller, R., & Chandiramani, S. (2022). Racial and gender differences of cervical vertebral maturation staging between African Americans versus Caucasian patients of various age groups: A retrospective study. *International Orthodontics*, *20*(1). https://doi.org/10.1016/j.ortho.2021.100606

Gupta, S., Jain, S., Gupta, P., & Deoskar, A. (2012). Determining skeletal maturation using insulin-like growth factor I (IGF-I) test. *Progress in Orthodontics*, *13*(3), 288–295. https://doi.org/10.1016/j.pio.2011.09.006

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.

Gyftopoulos, S., Lin, D., Knoll, F., Doshi, A. M., Rodrigues, T. C., & Recht, M. P. (2019). Artificial Intelligence in Musculoskeletal Imaging: Current Status and Future Directions. *American Journal of Roentgenology*, *213*(3), 506–513. https://doi.org/10.2214/AJR.19.21117

Hagg, U., & Pancherz, H. (1988). Dentofacial orthopaedics in relation to chronological age, growth period and skeletal development. An analysis of 72 male patients with Class II division 1 malocclusion treated with the Herbst appliance. *The European Journal of Orthodontics*, *10*(3), 169–176. https://doi.org/10.1093/ejo/10.3.169

Hägg, U., & Taranger, J. (1982). Maturation indicators and the pubertal growth spurt. *American Journal of Orthodontics*, *82*(4), 299–309. https://doi.org/10.1016/0002-9416(82)90464-X

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hassel, B., & Farman, A. G. (1995). Skeletal maturation evaluation using cervical vertebrae. *American Journal of Orthodontics and Dentofacial Orthopedics*, *107*(1), 58–66. https://doi.org/10.1016/S0889-5406(95)70157-5

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Houston, W. J. B. (1980). Relationships between skeletal maturity estimated from hand-wrist radiographs and the timing of the adolescent growth spurt. In *European Journal of Orthodontics* (Vol. 2).

Houston, W. J. B., Miller, J. C., & Tanner, J. M. (1979). Prediction of the Timing of the Adolescent Growth Spurt from Ossification Events in Hand—Wrist Films. *British Journal of Orthodontics*, *6*(3), 145–152. https://doi.org/10.1179/bjo.6.3.145

Hung, M., Park, J., Hon, E. S., Bounsanga, J., Moazzami, S., Ruiz-Negrón, B., & Wang, D. (2020). Artificial intelligence in dentistry: Harnessing big data to predict oral

cancer survival. *World Journal of Clinical Oncology*, *11*(11), 918–934. https://doi.org/10.5306/wjco.v11.i11.918

Hussain, U., Ul Hassan, F., Kamran, M. A., Alnazeh, A. A., Ullah, I., & Alam, S. (2024). Inter-observer and intra-observer agreement of cervical vertebral maturation staging: A systematic review and meta-analysis. *International Orthodontics*, *22*(3), 100874. https://doi.org/10.1016/j.ortho.2024.100874

Jaqueira, L. M. F., Armond, M. C., Pereira, L. J., Alcântara, C. E. P. de, & Marques, L. S. (2010). Determining skeletal maturation stage using cervical vertebrae: evaluation of three diagnostic methods. *Brazilian Oral Research*, *24*(4), 433–437. https://doi.org/10.1590/s1806-83242010000400010

Jayaraman, J., Roberts, G. J., King, N. M., & Wong, H. M. (2012). Dental age assessment of southern Chinese using the United Kingdom Caucasian reference dataset. *Forensic Science International*, *216*(1–3), 68–72. https://doi.org/10.1016/j.forsciint.2011.08.019

Jeong, M. J., Lee, K. E., Chae, Y. K., Nam, O. H., Lee, H. S., & Choi, S. C. (2022). Correlations between skeletal maturity and dental calcification stages in Koreanchildren. *European Journal of Paediatric Dentistry*, *23*(2), 101–105. https://doi.org/10.23804/ejpd.2022.23.02.03

Jiang, F., Abdulqader, A. A., Yan, Y., Cheng, F., Xiang, T., Yu, J., Li, J., Qiu, Y., & Chen, X. (2025). Deep learning based quantitative cervical vertebral maturation analysis. *Head and Face Medicine*, *21*(1). https://doi.org/10.1186/s13005-025-00498-6

Jocher, G. C. A. Q. J. (2023). *YOLO by Ultralytics (Version 8.0.0) [Computer software]*. Ultralytics. https://github.com/ultralytics/ultralytics

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Jones, C. M., Buchlak, Q. D., Oakden-Rayner, L., Milne, M., Seah, J., Esmaili, N., & Hachey, B. (2021). Chest radiographs and machine learning – Past, present and future. In *Journal of Medical Imaging and Radiation Oncology* (Vol. 65, Issue 5, pp. 538–544). John Wiley and Sons Inc. https://doi.org/10.1111/1754-9485.13274

Kapila, S., Vora, S. R., Rengasamy Venugopalan, S., Elnagar, M. H., & Akyalcin, S. (2023). Connecting the dots towards precision orthodontics. *Orthodontics & Craniofacial Research*, *26*(S1), 8–19. https://doi.org/10.1111/ocr.12725

Kazimierczak, N., Kazimierczak, W., Serafin, Z., Nowicki, P., Nożewski, J., & Janiszewska-Olszowska, J. (2024). AI in Orthodontics: Revolutionizing Diagnostics and Treatment Planning—A Comprehensive Review. In *Journal of Clinical Medicine* (Vol. 13, Issue 2). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/jcm13020344

Kazimierczak, W., Jedliński, M., Issa, J., Kazimierczak, N., Janiszewska-Olszowska, J., Dyszkiewicz-Konwińska, M., Różyło-Kalinowska, I., Serafin, Z., & Orhan, K. (2024). Accuracy of Artificial Intelligence for Cervical Vertebral Maturation Assessment—A Systematic Review. In *Journal of Clinical Medicine* (Vol. 13, Issue 14). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/jcm13144047

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems (NeurIPS)*, 3146–3154. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, *17*(1), 195. https://doi.org/10.1186/s12916-019-1426-2

Khade, D. M., Bhad, W. A., Chavan, S. J., Muley, A., & Shekokar, S. (2023). Reliability of salivary biomarkers as skeletal maturity indicators: A systematic review. In *International Orthodontics* (Vol. 21, Issue 1). Elsevier Masson s.r.l. https://doi.org/10.1016/j.ortho.2022.100716

Khanagar, S. B., Al-ehaideb, A., Maganur, P. C., Vishwanathaiah, S., Patil, S., Baeshen, H. A., Sarode, S. C., & Bhandi, S. (2021). Developments, application, and performance of artificial intelligence in dentistry – A systematic review. In *Journal of Dental Sciences* (Vol. 16, Issue 1, pp. 508–522). Association for Dental Sciences of the Republic of China. https://doi.org/10.1016/j.jds.2020.06.019

Khanagar, S. B., Al-Ehaideb, A., Vishwanathaiah, S., Maganur, P. C., Patil, S., Naik, S., Baeshen, H. A., & Sarode, S. S. (2021). Scope and performance of artificial

intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - A systematic review. In *Journal of Dental Sciences* (Vol. 16, Issue 1, pp. 482–492). Association for Dental Sciences of the Republic of China. https://doi.org/10.1016/j.jds.2020.05.022

Khazaei, M., Mollabashi, V., Khotanlou, H., & Farhadian, M. (2023). Automatic determination of pubertal growth spurts based on the cervical vertebral maturation staging using deep convolutional neural networks. *Journal of the World Federation of Orthodontists*, *12*(2), 56–63. https://doi.org/10.1016/j.ejwf.2023.02.003

Kim, D. W., Kim, J., Kim, T., Kim, T., Kim, Y. J., Song, I. S., Ahn, B., Choo, J., & Lee, D. Y. (2021). Prediction of hand-wrist maturation stages based on cervical vertebrae images using artificial intelligence. *Orthodontics and Craniofacial Research*, *24*(S2), 68–75. https://doi.org/10.1111/ocr.12514

Kim, S. H., Choi, Y. K., Shin, S. M., Choi, Y. S., Yamaguchi, T., Takahashi, M., Maki, K., Park, S. B., & Kim, Y. Il. (2017). The estimation of skeletal maturity of patients with cleft lip and palate using statistical shape analysis: A preliminary study. *Dentomaxillofacial Radiology*, *46*(5). https://doi.org/10.1259/dmfr.20160491

Kohli, R. K., Arora, M., & Puri, P. (2024). Assessment of Cervical Vertebrae Maturation Stages and the Accuracy of Age Estimation Methods in North Indian Population with Digital Lateral Cephalogram. *Journal of Indian Academy of Forensic Medicine*, *46*(4), 448–454. https://doi.org/10.1177/09710973251327636

Kök, H., Acilar, A. M., & İzgi, M. S. (2019). Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics. *Progress in Orthodontics*, *20*(1), 41–51. https://doi.org/10.1186/s40510-019-0295-8

Koshy, S., & Tandon, S. (1998). Dental age assessment: The applicability of Demirjian's method in South Indian children. In *Forensic Science International* (Vol. 94).

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, *64*(1), 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002

Kunz, F., Stellzig-Eisenhauer, A., Zeman, F., & Boldt, J. (2020). Artificial intelligence in orthodontics: Evaluation of a fully automated cephalometric analysis using a customized convolutional neural network. *Journal of Orofacial Orthopedics*, *81*(1), 52–68. https://doi.org/10.1007/s00056-019-00203-8

Lamparski, D. G. (1972). *Skeletal Age Assessment Utilizing Cervical Vertebrae*. University of Pittsburgh.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lee, J. H., Yu, H. J., Kim, M. J., Kim, J. W., & Choi, J. (2020). Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. *BMC Oral Health*, *20*(1). https://doi.org/10.1186/s12903-020-01256-7

Lee, J. M., Moon, J. H., Park, J. A., Kim, J. H., & Lee, S. J. (2024). Factors influencing the development of artificial intelligence in orthodontics. *Orthodontics and Craniofacial Research*. https://doi.org/10.1111/ocr.12806

Lee, M. K., Allareddy, V., Rampa, S., Elnagar, M. H., Oubaidin, M., Yadav, S., & Rengasamy Venugopalan, S. (2024). Applications and challenges of implementing artificial intelligence in orthodontics: A primer for orthodontists. *Seminars in Orthodontics*, *30*(1), 72–76. https://doi.org/10.1053/j.sodo.2024.01.005

Li, H., Chen, Y., Wang, Q., Gong, X., Lei, Y., Tian, J., & Gao, X. (2022). Convolutional neural network-based automatic cervical vertebral maturation classification method. *Dentomaxillofacial Radiology*, *51*(6). https://doi.org/10.1259/dmfr.20220070

Li, H., Li, H., Yuan, L., Liu, C., Xiao, S., Liu, Z., Zhou, G., Dong, T., Ouyang, N., Liu, L., Ma, C., Feng, Y., Zheng, Y., Xia, L., & Fang, B. (2023). The psc-CVM assessment system: A three-stage type system for CVM assessment based on deep learning. *BMC Oral Health*, *23*(1), 557–568. https://doi.org/10.1186/s12903-023-03266-7

Li, S., Guo, Z., Lin, J., & Ying, S. (2022). Artificial Intelligence for Classifying and Archiving Orthodontic Images. *BioMed Research International*, *2022*. https://doi.org/10.1155/2022/1473977

Li, Z., & Wang, L. (2025). Multi-task reinforcement learning and explainable AI-Driven platform for personalized planning and clinical decision support in

orthodontic-orthognathic treatment. *Scientific Reports*, *15*(1), 24502. https://doi.org/10.1038/s41598-025-09236-z

Liao, N., Dai, J., Tang, Y., Zhong, Q., & Mo, S. (2022). iCVM: An Interpretable Deep Learning Model for CVM Assessment Under Label Uncertainty. *IEEE Journal of Biomedical and Health Informatics*, *26*(8), 4325–4334. https://doi.org/10.1109/JBHI.2022.3179619

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. In *Medical Image Analysis* (Vol. 42, pp. 60–88). Elsevier B.V. https://doi.org/10.1016/j.media.2017.07.005

Liu, H., Feng, L., & Wang, L. (2023). Diagnostic value of cervical vertebral maturation stages for midpalatal suture maturation assessment: a study in the Chinese population. *BMC Oral Health*, *23*(1), 504. https://doi.org/10.1186/s12903-023-03220-7

Liu, J., Chen, Y., Li, S., Zhao, Z., & Wu, Z. (2021). Machine learning in orthodontics: Challenges and perspectives. In *Advances in Clinical and Experimental Medicine* (Vol. 30, Issue 10). Wroclaw University of Medicine. https://doi.org/10.17219/acem/138702

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, *1*(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2

Machado, V., Botelho, J., Mascarenhas, P., Mendes, J. J., & Delgado, A. (2020). A systematic review and meta-analysis on Bolton's ratios: Normal occlusion and malocclusion. *Journal of Orthodontics*, *47*(1), 7–29. https://doi.org/10.1177/1465312519886322

Magalhães, M. I., Machado, V., Mascarenhas, P., Botelho, J., Mendes, J. J., & Delgado, A. S. (2022). Chronological age range estimation of cervical vertebral maturation using Baccetti method: a systematic review and meta-analysis. *European Journal of Orthodontics*, *44*(5), 548–555. https://doi.org/10.1093/ejo/cjac009

Magat, G., & Ozcan, S. (2022). Assessment of maturation stages and the accuracy of age estimation methods in a Turkish population: A comparative study. *Imaging Science in Dentistry*, *52*. https://doi.org/10.5624/ISD.20210231

Makaremi, M., Lacaule, C., & Mohammad-Djafari, A. (2019). Deep learning and artificial intelligence for the determination of the cervical vertebra maturation degree from lateral radiography. *Entropy*, *21*(12), 1–24. https://doi.org/10.3390/e21121222

Malamud, D. (2011). Saliva as a Diagnostic Fluid. *Dental Clinics of North America*, *55*(1), 159–178. https://doi.org/10.1016/j.cden.2010.08.004

Mandall, N., Cousley, R., DiBiase, A., Dyer, F., Littlewood, S., Mattick, R., Nute, S. J., Doherty, B., Stivaros, N., McDowall, R., Shargill, I., & Worthington, H. V. (2016a). Early class III protraction facemask treatment reduces the need for orthognathic surgery: a multi-centre, two-arm parallel randomized, controlled trial. *Journal of Orthodontics*, *43*(3), 164–175. https://doi.org/10.1080/14653125.2016.1201302

Mandall, N., Cousley, R., DiBiase, A., Dyer, F., Littlewood, S., Mattick, R., Nute, S. J., Doherty, B., Stivaros, N., McDowall, R., Shargill, I., & Worthington, H. V. (2016b). Early class III protraction facemask treatment reduces the need for orthognathic surgery: a multi-centre, two-arm parallel randomized, controlled trial. *Journal of Orthodontics*, *43*(3), 164–175. https://doi.org/10.1080/14653125.2016.1201302

Mandall, N., Littlewood, S., Shah, J., Watkinson, S., Glossop, S., Dugdale, C., Curran, E., & Walsh, T. (2022). The effect of treatment timing on clinical and psychological outcomes with twin block therapy: A multicentre two-arm parallel randomised controlled trial. *Journal of Orthodontics*, 146531252211178. https://doi.org/10.1177/14653125221117845

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Masoud, M., Masoud, I., Kent, R. L., Gowharji, N., & Cohen, L. E. (2008). Assessing skeletal maturity by using blood spot insulin-like growth factor I (IGF-I) testing. *American Journal of Orthodontics and Dentofacial Orthopedics*, *134*(2), 209–216. https://doi.org/10.1016/j.ajodo.2006.09.063

Mathew, R., Palatinus, S., Padala, S., Alshehri, A., Awadh, W., Bhandi, S., Thomas, J., & Patil, S. (2022). Neural networks for classification of cervical vertebrae maturation: a systematic review. *The Angle Orthodontist*, *92*(6), 796–804. https://doi.org/10.2319/031022-210.1

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

McNamara, J. A., & Franchi, L. (2018). The cervical vertebral maturation method: A user's guide. *Angle Orthodontist*, *88*(2), 133–143. https://doi.org/10.2319/111517-787.1

Mellion, Z. J., Behrents, R. G., & Johnston, L. E. (2013). The pattern of facial skeletal growth and its relationship to various common indexes of maturation. *American Journal of Orthodontics and Dentofacial Orthopedics*, *143*(6), 845–854. https://doi.org/10.1016/j.ajodo.2013.01.019

Miltenburg Caspersen, L., & Sonnesen, L. (2020). Secular trend of the skeletal maturation in relation to peak height velocity-a comparison between two groups of children born 1969-1973 and 1996-2000. In *European Journal of Orthodontics* (Vol. 42, Issue 6, pp. 612–618). Oxford University Press. https://doi.org/10.1093/ejo/cjz098

Mohammad, N., Muad, A. M., Ahmad, R., & Yusof, M. Y. P. M. (2022). Accuracy of advanced deep learning with tensorflow and keras for classifying teeth developmental stages in digital panoramic imaging. *BMC Medical Imaging*, *22*(66), 1–13. https://doi.org/10.1186/s12880-022-00794-6

Mohammad-Rahimi, H., Motamadian, S. R., Nadimi, M., Hassanzadeh-Samani, S., Minabi, M. A. S., Mahmoudinia, E., Lee, V. Y., & Rohban, M. H. (2022). Deep learning for the classification of cervical maturation degree and pubertal growth spurts: A pilot study. *Korean Journal of Orthodontics*, *52*(2), 112–122. https://doi.org/10.4041/kjod.2022.52.2.112

Mohammad-Rahimi, H., Nadimi, M., Rohban, M. H., Shamsoddin, E., Lee, V. Y., & Motamedian, S. R. (2021). Machine learning and orthodontics, current trends and the future opportunities: A scoping review. In *American Journal of Orthodontics and Dentofacial Orthopedics* (Vol. 160, Issue 2, pp. 170-192.e4). Mosby Inc. https://doi.org/10.1016/j.ajodo.2021.02.013

Mohammed, M. H., Omer, Z. Q., Aziz, B. B., Abdulkareem, J. F., Mahmood, T. M. A., Kareem, F. A., & Mohammad, D. N. (2024). Convolutional Neural Network-

Based Deep Learning Methods for Skeletal Growth Prediction in Dental Patients. *Journal of Imaging*, *10*(11). https://doi.org/10.3390/jimaging10110278

Morris, K. M., Fields, H. W., Beck, F. M., & Kim, D. G. (2019). Diagnostic testing of cervical vertebral maturation staging: An independent assessment. *American Journal of Orthodontics and Dentofacial Orthopedics*, *156*(5), 626–632. https://doi.org/10.1016/j.ajodo.2018.11.016

Nestman, T. S., Marshall, S. D., Qian, F., Holton, N., Franciscus, R. G., & Southard, T. E. (2011). Cervical vertebrae maturation method morphologic criteria: Poor reproducibility. *American Journal of Orthodontics and Dentofacial Orthopedics*, *140*(2), 182–188. https://doi.org/10.1016/j.ajodo.2011.04.013

Ngan, P., & Moon, W. (2015). Evolution of Class III treatment in orthodontics. *American Journal of Orthodontics and Dentofacial Orthopedics*, *148*(1), 22–36. https://doi.org/10.1016/j.ajodo.2015.04.012

Norman, N. H., Mohd Rosli, M., Mohammed Al-Jaf, N., Mohammad, N., & Mohd Yusof, M. Y. P. (2025). Automated cervical vertebral maturation staging using deep learning: Enhancing accuracy through random oversampling and memory optimization. *APOS Trends in Orthodontics*, *0*, 1. https://doi.org/10.25259/APOS_322_2024

Norman, N. H., Rosli, M. M., Al-Jaf, N. M., Mohammad, N., Azizan, A., & Yusof, M. Y. P. M. (2025). Integration of artificial intelligence in orthodontic imaging: A bibliometric analysis of research trends and applications. *Imaging Science in Dentistry*, *55*(2), 151. https://doi.org/10.5624/isd.20240237

Nykänen, R., Espeland, L., Kvaal, S. I., & Krogstad, O. (1998). Validity of the Demirjian method for dental age estimation when applied to Norwegian children. *Acta Odontologica Scandinavica*, *56*(4), 238–244. https://doi.org/10.1080/00016359850142862

Nyström, M., Aine, L., Peck, L., Haavikko, K., & Kataja, M. (2000). Dental maturity in Finns and the problem of missing teeth. *Acta Odontologica Scandinavica*, *58*(2), 49–56. https://doi.org/10.1080/000163500429280

O'Brien, K., Wright, J., Conboy, F., Appelbe, P., Davies, L., Connolly, I., Mitchell, L., Littlewood, S., Mandall, N., Lewis, D., Sandler, J., Hammond, M., Chadwick, S., O'Neill, J., McDade, C., Oskouei, M., Thiruvenkatachari, B., Read, M., Robinson, S., … Worthington, H. (2009). Early treatment for Class II Division 1 malocclusion with the Twin-block appliance: A multi-center, randomized,

controlled trial. *American Journal of Orthodontics and Dentofacial Orthopedics*, *135*(5), 573–579. https://doi.org/10.1016/j.ajodo.2007.10.042

O'Brien, K., Wright, J., Conboy, F., Sanjie, Y. W., Mandall, N., Chadwick, S., Connolly, I., Cook, P., Birnie, D., Hammond, M., Harradine, N., Lewis, D., McDade, C., Mitchell, L., Murray, A., O'Neill, J., Read, M., Robinson, S., Roberts-Harry, D., … Shaw, I. (2003). Effectiveness of early orthodontic treatment with the Twin-block appliance: A multicenter, randomized, controlled trial. Part 1: Dental and skeletal effects. *American Journal of Orthodontics and Dentofacial Orthopedics*, *124*(3), 234–243. https://doi.org/10.1016/S0889-5406(03)00352-4

Pandey, H., Tripathi, V., Pathak, H., Choudhary, S., & Parchake, M. (2019). Age estimation and comparison by dental and skeletal maturity in the age range of 9–18 years in the Mumbai region. *Journal of Forensic Dental Sciences*, *11*(3), 142. https://doi.org/10.4103/jfo.jfds_90_19

Patel, P., Chaudhary, A., Dudhia, B., Bhatia, P., Jani, Y., & Soni, N. (2015). Accuracy of two dental and one skeletal age estimation methods in 6-16 year old Gujarati children. *Journal of Forensic Dental Sciences*, *7*(1), 18. https://doi.org/10.4103/0975-1475.150298

Perillo, L., Masucci, C., Ferro, F., Apicella, D., & Baccetti, T. (2010). Prevalence of Orthodontic Treatment Need in Southern Italian Schoolchildren. *The European Journal of Orthodontics*, *32*(1), 49–53. https://doi.org/10.1093/ejo/cjp050

Perinetti, G., Baccetti, T., Contardo, L., & Di Lenarda, R. (2011). Gingival crevicular fluid alkaline phosphatase activity as a non-invasive biomarker of skeletal maturation. *Orthodontics and Craniofacial Research*, *14*(1), 44–50. https://doi.org/10.1111/j.1601-6343.2010.01506.x

Perinetti, G., Braga, C., Contardo, L., & Primozic, J. (2020). Cervical vertebral maturation: Are postpubertal stages attained in all subjects? *American Journal of Orthodontics and Dentofacial Orthopedics*, *157*(3), 305–312. https://doi.org/10.1016/j.ajodo.2019.03.026

Perinetti, G., Caprioglio, A., & Contardo, L. (2014). Visual assessment of the cervical vertebral maturation stages a study of diagnostic accuracy and repeatability. *Angle Orthodontist*, *84*(6), 951–956. https://doi.org/10.2319/120913-906.1

Perinetti, G., & Contardo, L. (2017). Reliability of Growth Indicators and Efficiency of Functional Treatment for Skeletal Class II Malocclusion: Current Evidence and

Controversies. In *BioMed Research International* (Vol. 2017, pp. 1–19). Hindawi Limited. https://doi.org/10.1155/2017/1367691

Perinetti, G., Contardo, L., Gabrieli, P., Baccetti, T., & Di Lenarda, R. (2012). Diagnostic performance of dental maturity for identification of skeletal maturation phase. *European Journal of Orthodontics*, *34*(4), 487–492. https://doi.org/10.1093/ejo/cjr027

Perinetti, G., Di Lenarda, R., & Contardo, L. (2013). Diagnostic performance of combined canine and second molar maturity for identification of growth phase. *Progress in Orthodontics*, *14*(1), 1. https://doi.org/10.1186/2196-1042-14-1

Perinetti, G., Primozic, J., Franchi, L., & Contardo, L. (2016). Cervical vertebral maturation method: growth timing versus growth amount. *The European Journal of Orthodontics*, *38*(1), 111–112. https://doi.org/10.1093/ejo/cjv037

Perinetti, G., Primozic, J., Sharma, B., Cioffi, I., & Contardo, L. (2018). Cervical vertebral maturation method and mandibular growth peak: A longitudinal study of diagnostic reliability. *European Journal of Orthodontics*, *40*(6), 666–672. https://doi.org/10.1093/ejo/cjy018

Pirri, S., Lorenzoni, V., & Turchetti, G. (2020). Scoping review and bibliometric analysis of Big Data applications for Medication adherence: An explorative methodological study to enhance consistency in literature. *BMC Health Services Research*, *20*(1). https://doi.org/10.1186/s12913-020-05544-4

Polizzi, A., & Leonardi, R. (2024). Automatic cephalometric landmark identification with artificial intelligence: An umbrella review of systematic reviews. In *Journal of Dentistry* (Vol. 146). Elsevier Ltd. https://doi.org/10.1016/j.jdent.2024.105056

Powers, D. M. W. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*.

Prasad, J., Mallikarjunaiah, D., Shetty, A., Gandedkar, N., Chikkamuniswamy, A., & Shivashankar, P. (2022). Machine Learning Predictive Model as Clinical Decision Support System in Orthodontic Treatment Planning. *Dentistry Journal*, *11*(1), 1. https://doi.org/10.3390/dj11010001

Pratiwi, D., Kusumadewy, W., Adiatman, M., Suniarti, D. F., Soedarsono, N., & Soegiharto, B. M. (2025). The reproducibility of cervical vertebrae maturation: A cross-sectional comparative study between two methods. *F1000Research*, *14*, 166. https://doi.org/10.12688/f1000research.160533.1

Predko-Engel, A., Kaminek, M., Langova, K., Kowalski, P., & Fudalej, P. S. (2015). Reliability of the cervical vertebrae maturation (CVM) method. *Bratislava Medical Journal*, *116*(4), 222–226. https://doi.org/10.4149/BLL_2015_043

Proffit, W. R. ., Fields, H. W. ., Larson, B. E. ., & Sarver, D. M. . (2018). *Contemporary Orthodontics* (6th ed.). Elsevier.

Proffit, W. R., Turvey, T. A., & Phillips, C. (2007). The hierarchy of stability and predictability in orthognathic surgery with rigid fixation: an update and extension. *Head & Face Medicine*, *3*(1), 21. https://doi.org/10.1186/1746-160X-3-21

Pyle, S., & Greulich, W. (1959). *Radiographic Atlas of Skeletal Development of Hand and Wrist* (2nd ed.). Stanford University Press.

Rainey, B.-J., Burnside, G., & Harrison, J. E. (2016). Reliability of cervical vertebral maturation staging. *American Journal of Orthodontics and Dentofacial Orthopedics*, *150*(1), 98–104. https://doi.org/10.1016/j.ajodo.2015.12.013

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., Patel, B. N., Yeom, K. W., Shpanskaya, K., Blankenberg, F. G., Seekins, J., Amrhein, T. J., Mong, D. A., Halabi, S. S., Zucker, E. J., … Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*, *15*(11). https://doi.org/10.1371/journal.pmed.1002686

Ratitch, B., Trigg, A., Majumder, M., Vlajnic, V., Rethemeier, N., & Nkulikiyinka, R. (2023). Clinical Validation of Novel Digital Measures: Statistical Methods for Reliability Evaluation. *Digital Biomarkers*, 74–91. https://doi.org/10.1159/000531054

Robins, S. P. (1994). Biochemical markers for assessing skeletal growth. *European Journal of Clinical Nutrition*, *48 Suppl 1*, S199-209.

San Román, P., Palma, J. C., Oteo, M. D., & Nevado, E. (2002). Skeletal maturation determined by cervical vertebrae development. *European Journal of Orthodontics*, *24*(3), 303–311. https://doi.org/10.1093/ejo/24.3.303

Santana, L. G., Gatti-Reis, L., Paiva, S. M., Ramos-Jorge, M. L., & Marques, L. S. (2025). Patient-centered factors associated with orthodontic treatment success: a scoping review. *Brazilian Oral Research*, *39*. https://doi.org/10.1590/1807-3107bor-2025.vol39.020

Santiago, R. C., De Miranda Costa, L. F., Vitral, R. W. F., Fraga, M. R., Bolognese, A. M., & Maia, L. C. (2012). Cervical vertebral maturation as a biologic indicator of skeletal maturity: A systematic review. In *Angle Orthodontist* (Vol. 82, Issue 6, pp. 1123–1131). https://doi.org/10.2319/103111-673.1

Schlégl, Á. T., O'Sullivan, I., Varga, P., Than, P., & Vermes, C. (2022). Alternative methods for skeletal maturity estimation with the EOS scanner—Experience from 934 patients. *PLoS ONE*, *17*(5 May). https://doi.org/10.1371/journal.pone.0267668

Schoretsaniti, L., Mitsea, A., Karayianni, K., & Sifakakis, I. (2021). Cervical Vertebral Maturation Method: Reproducibility and Efficiency of Chronological Age Estimation. *Applied Sciences*, *11*(7), 3160. https://doi.org/10.3390/app11073160

Schwendicke, F., Golla, T., Dreher, M., & Krois, J. (2019). Convolutional neural networks for dental image diagnostics: A scoping review. In *Journal of Dentistry* (Vol. 91). Elsevier Ltd. https://doi.org/10.1016/j.jdent.2019.103226

Sella Tunis, T., Masarwa, M., Finkelstein, T., Grinstein Koren, O., Kats, L., Manor, Y., Shpack, N., & Reiter, S. (2024). The reliability of a modified three-stage cervical vertebrae maturation method for estimating skeletal growth in males and females. *BMC Oral Health*, *24*(1). https://doi.org/10.1186/s12903-024-05028-5

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Seo, H., Hwang, J., Jeong, T., & Shin, J. (2021). Comparison of deep learning models for cervical vertebral maturation stage classification on lateral cephalometric radiographs. *Journal of Clinical Medicine*, *10*(16), 1–11. https://doi.org/10.3390/jcm10163591

Shahbazi, S., Esmaeili, S., Kavousinejad, S., Younessian, F., & Behnaz, M. (2025). Efficacy of artificial intelligence in radiographic dental age estimation of patients undergoing dental maturation: A systematic review and meta-analysis. *International Orthodontics*, *23*(4), 101010. https://doi.org/10.1016/j.ortho.2025.101010

Shan, T., Tay, F. R., & Gu, L. (2021). Application of Artificial Intelligence in Dentistry. In *Journal of Dental Research* (Vol. 100, Issue 3, pp. 232–244). SAGE Publications Inc. https://doi.org/10.1177/0022034520969115

Shaw, W. C., O'Brien, K. D., & Richmond, S. (1991). Quality Control In Orthodontics: Factors Influencing The Receipt of Orthodontic Treatment. *British Dental Journal*, *170*(2), 66–68. https://doi.org/10.1038/sj.bdj.4807421

Shim, J. J., Heo, G., & Lagravère, M. O. (2012). Assessment of skeletal maturation based on cervical vertebrae in CBCT. *International Orthodontics*, *10*(4), 351–362. https://doi.org/10.1016/j.ortho.2012.09.005

Shoari, S. A., Sadrolashrafi, S. V., Sohrabi, A., Afrouzian, R., Ebrahimi, P., Kouhsoltani, M., & Soltani, M. K. (2024). Estimating mandibular growth stage based on cervical vertebral maturation in lateral cephalometric radiographs using artificial intelligence. *Progress in Orthodontics*, *25*(28), 1–12. https://doi.org/10.1186/s40510-024-00527-1

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*, *320*(21), 2199. https://doi.org/10.1001/jama.2018.17163

Sierra, A. M. (1987). Assessment of dental and skeletal maturity. A new approach. *The Angle Orthodontist*, *57*(3), 194–208. https://doi.org/10.1043/0003-3219(1987)057<0194:AODASM>2.0.CO;2

Soegiharto, B. M., Cunningham, S. J., & Moles, D. R. (2008). Skeletal maturation in Indonesian and white children assessed with hand-wrist and cervical vertebrae methods. *American Journal of Orthodontics and Dentofacial Orthopedics*, *134*(2), 217–226. https://doi.org/10.1016/j.ajodo.2006.07.037

Soegiharto, B. M., Moles, D. R., & Cunningham, S. J. (2008). Discriminatory ability of the skeletal maturation index and the cervical vertebrae maturation index in detecting peak pubertal growth in Indonesian and white subjects with receiver operating characteristics analysis. *American Journal of Orthodontics and Dentofacial Orthopedics*, *134*(2), 227–237. https://doi.org/10.1016/j.ajodo.2006.09.062

Sohrabi, A., Babay Ahari, S., Moslemzadeh, H., Rafighi, A., & Aghazadeh, Z. (2016). The reliability of clinical decisions based on the cervical vertebrae maturation staging method. *European Journal of Orthodontics*, *38*(1), 8–12. https://doi.org/10.1093/ejo/cjv030

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Stahl, F., Baccetti, T., Franchi, L., & McNamara, J. A. (2008). Longitudinal growth changes in untreated subjects with Class II Division 1 malocclusion. *American Journal of Orthodontics and Dentofacial Orthopedics*, *134*(1), 125–137. https://doi.org/10.1016/j.ajodo.2006.06.028

Szemraj, A., Wojtaszek-Słomińska, A., & Racka-Pilszak, B. (2018). Is the cervical vertebral maturation (CVM) method effective enough to replace the hand-wrist maturation (HWM) method in determining skeletal maturation?—A systematic review. *European Journal of Radiology*, *102*, 125–128. https://doi.org/10.1016/j.ejrad.2018.03.012

Szemraj-Folmer, A., Wojtaszek-Słomińska, A., Racka-Pilszak, B., & Kuc-Michalska, M. (2021). Assessment of the duration of the pubertal growth spurt in patients with skeletal open bite: A cross-sectional study. *Journal of Orofacial Orthopedics*, *82*(2), 92–98. https://doi.org/10.1007/s00056-020-00262-2

Tang, A., Tam, R., Cadrin-Chênevert, A., Guest, W., Chong, J., Barfett, J., Chepelev, L., Cairns, R., Mitchell, J. R., Cicero, M. D., Poudrette, M. G., Jaremko, J. L., Reinhold, C., Gallix, B., Gray, B., Geis, R., O'Connell, T., Babyn, P., Koff, D., … Shabana, W. (2018). Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. In *Canadian Association of Radiologists Journal* (Vol. 69, Issue 2, pp. 120–135). Canadian Medical Association. https://doi.org/10.1016/j.carj.2018.02.002

Tejani, A. S., Klontzas, M. E., Gatti, A. A., Mongan, J. T., Moy, L., Park, S. H., Kahn, C. E., Abbara, S., Afat, S., Anazodo, U. C., Andreychenko, A., Asselbergs, F. W., Badano, A., Baessler, B., Bold, B., Bisdas, S., Brismar, T. B., Cacciamani, G. E., Carrino, J. A., … Zins, M. (2024). Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence*, *6*(4). https://doi.org/10.1148/ryai.240300

Tekın, A., & Cesur Aydın, K. (2020). Comparative determination of skeletal maturity by hand–wrist radiograph, cephalometric radiograph and cone beam computed tomography. *Oral Radiology*, *36*(4), 327–336. https://doi.org/10.1007/s11282-019-00408-y

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. . *4th Machine Learning for Healthcare Conference*, 359–380. https://proceedings.mlr.press/v106/tonekaboni19a.html

Topol, E. J. (2019). *Deep Medicine : How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books.

Trehan, M., & Patil, C. (2021). Evaluation of alkaline phosphatase as skeletal maturity indicator in gingival crevicular fluid. *International Journal of Clinical Pediatric Dentistry*, *14*(4), 512–517. https://doi.org/10.5005/jp-journals-10005-1996

Tripathi, T., Gupta, P., Rai, P., Sharma, J., Gupta, V. K., & Singh, N. (2017). Osteocalcin and serum insulin-like growth factor-1 as biochemical skeletal maturity indicators. *Progress in Orthodontics*, *18*(1). https://doi.org/10.1186/s40510-017-0184-y

Tripathi, T., Gupta, P., Sharma, J., Rai, P., Gupta, V. K., & Singh, N. (2018). Bone-specific alkaline phosphatase–a potential biomarker for skeletal growth assessment. *Journal of Orthodontics*, *45*(1), 4–10. https://doi.org/10.1080/14653125.2017.1416571

Tsirtsakis, P., Zacharis, G., Maraslidis, G. S., & Fragulis, G. F. (2025). Deep learning for object recognition: A comprehensive review of models and algorithms. *International Journal of Cognitive Computing in Engineering*, *6*, 298–312. https://doi.org/10.1016/j.ijcce.2025.01.004

Tsui, K.-L., Wong, Z. S.-Y., Goldsman, D., & Edesess, M. (2013). Tracking Infectious Disease Spread for Global Pandemic Containment. *IEEE Intelligent Systems*, *28*(6), 60–64. https://doi.org/10.1109/MIS.2013.149

Tulloch, J. F. C., Proffit, W. R., & Phillips, C. (2004). Outcomes in a 2-phase randomized clinical trial of early class II treatment. *American Journal of Orthodontics and Dentofacial Orthopedics*, *125*(6), 657–667. https://doi.org/10.1016/j.ajodo.2004.02.008

Tzutalin. (2015). *LabelImg*. GitHub. https://github.com/tzutalin/labelImg

Ueda, A., Tussie, C., Kim, S., Kuwajima, Y., Matsumoto, S., Kim, G., Satoh, K., & Nagai, S. (2023). Classification of Maxillofacial Morphology by Artificial Intelligence Using Cephalometric Analysis Measurements. *Diagnostics*, *13*(13), 2134. https://doi.org/10.3390/diagnostics13132134

Uysal, T., Ramoglu, S. I., Basciftci, F. A., & Sari, Z. (2006). Chronologic age and skeletal maturation of the cervical vertebrae and hand-wrist: Is there a relationship? *American Journal of Orthodontics and Dentofacial Orthopedics*, *130*(5), 622–628. https://doi.org/10.1016/j.ajodo.2005.01.031

Valentin, J. . (2007). *The 2007 Recommendations of the International Commission on Radiological Protection*. Published for the International Commission on Radiological Protection by Elsevier.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine*, *37*(5), 360–363.

Wayne, M., & Ellner, P. (2010). A comparison of the robustness of reliability growth assessment techniques. *2010 Proceedings - Annual Reliability and Maintainability Symposium (RAMS)*, 1–6. https://doi.org/10.1109/RAMS.2010.5448064

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. In *Journal of Strength and Conditioning Research* (Vol. 19, Issue 1, pp. 231–240). https://doi.org/10.1519/15184.1

White, S. C. ., & Pharoah, M. J. . (2014). *Oral Radiology : Principles and Interpretation*. Elsevier.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, *25*(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Willems, G., Van Olmen, A., Spiessens, B., & Carels, C. (2001). Dental age estimation in Belgian children: Demirjian's technique revisited. *Journal of Forensic Sciences*, *46*(4), 893–895.

World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health : WHO Guidance*. World Health Organization.

Wu, J. T., Wong, K. C. L., Gur, Y., Ansari, N., Karargyris, A., Sharma, A., Morris, M., Saboury, B., Ahmad, H., Boyko, O., Syed, A., Jadhav, A., Wang, H., Pillai, A., Kashyap, S., Moradi, M., & Syeda-Mahmood, T. (2020). Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. *JAMA Network Open*, *3*(10). https://doi.org/10.1001/jamanetworkopen.2020.22779

Yang, Y. M., Lee, J., Kim, Y. I., Cho, B. H., & Park, S. B. (2014). Axial cervical vertebrae-based multivariate regression model for the estimation of skeletal-maturation status. *Orthodontics and Craniofacial Research*, *17*(3), 187–196. https://doi.org/10.1111/ocr.12045

Yousefi, F., Mohammadi, Y., Ahmadvand, M., & Razaghi, P. (2023). Dental age estimation using cone-beam computed tomography: A systematic review and meta-analysis. *Imaging Science in Dentistry*, *53*(2), 91. https://doi.org/10.5624/isd.20221226

Yousefirizi, F., Pierre Decazes, Amyar, A., Ruan, S., Saboury, B., & Rahmim, A. (2022). AI-Based Detection, Classification and Prediction/Prognosis in Medical Imaging. *PET Clinics*, *17*(1), 183–212. https://doi.org/10.1016/j.cpet.2021.09.010

Zhang, Y., Lu, Z., Zhou, J., Sun, Y., Yi, W., Wang, J., Du, T., Li, D., Zhao, X., Xu, Y., Li, C., & Qi, K. (2025). CDSNet: An automated method for assessing growth stages from various anatomical regions in lateral cephalograms based on deep learning. *Journal of the World Federation of Orthodontists*, *14*(3), 154–160. https://doi.org/10.1016/j.ejwf.2024.09.007

Zhang, Y., Shu, S., Gu, Q., Liu, Z., Zhu, Z., Qiu, Y., & Bao, H. (2020). Cervical vertebral maturation (CVM) stage as a supplementary indicator for the assessment of peak height velocity (PHV) in adolescent idiopathic scoliosis (AIS). *Quantitative Imaging in Medicine and Surgery*, *10*(1), 96–105. https://doi.org/10.21037/qims.2019.11.07

Zhao, X. G., Lin, J., Jiang, J. H., Wang, Q., & Hong, S. N. G. (2012). Validity and reliability of a method for assessment of cervical vertebral maturation. *Angle Orthodontist*, *82*(2), 229–234. https://doi.org/10.2319/051511-333.1

Zhou, J., Zhou, H., Pu, L., Gao, Y., Tang, Z., Yang, Y., You, M., Yang, Z., Lai, W., & Long, H. (2021). Development of an artificial intelligence system for the automatic evaluation of cervical vertebral maturation status. *Diagnostics*, *11*(12). https://doi.org/10.3390/diagnostics11122200

APPENDICES

APPENDIX A

List of Publications

1. Noraina Hafizan Norman, Marshima Mohd Rosli, Nagham Mohammed Al-Jaf, Norhasmira Mohammad, Mohd Yusmiaidil Putera Mohd Yusof. *Automated Cervical Maturation Staging using Deep Learning: Enhancing Accuracy Through Random Oversampling and Memory Optimization.* **APOS Trends in Orthodontics.** 2025 June: Advanced Online Publication. DOI:10.25259/APOS_322_2024 ***(SCOPUS-indexed)***.

2. Noraina Hafizan Norman, Marshima Mohd Rosli, Nagham Mohammed Al-Jaf, Norhasmira Mohammad, Azliyana Azizan, Mohd Yusmiaidil Putera Mohd Yusof. *Integration of Artificial Intelligence in Orthodontic Imaging: A Bibliometric Analysis of Research Trends and Applications.* **Imaging Science in Dentistry.** 2025 June: 55 (2): 151-64. DOI: 10.5624/isd.20240237 ***(Web of Science Q2/*** *Emerging Sources Citation Index (ESCI)* ***SCOPUS-indexed)***

3. Noraina Hafizan Norman, Marshima Mohd Rosli, Nagham Mohammed Al-Jaf, Norhasmira Mohammad, Azliyana Azizan, Mohd Yusmiaidil Putera Mohd Yusof. *A Bibliometric Analysis of Artificial Intelligence in Orthodontics: Current Status, Development, and Future Research Directions.* EOS 2024 Abstracts, ***European Journal of Orthodontics***. 2025 April: 47(2): cjae081, DOI: 10.1093/ejo/cjae081.536 ***(Web of Science/ Science Citation Index Expanded (SCIE)/ SCOPUS-indexed)***

Others

4. Al Imran Shahrul, Nabilla Mohd Shukor, Noraina Hafizan Norman. *Does the Camera Type Affect the Quality of Orthodontic Photographs? Mirrorless and Smartphone Cameras Versus Digital Single Lens Reflex (DSLR) Camera.* **Journal of Orthodontics.** 2025 Aug:14653125251358837. DOI: 10.1177/14653125251358837. Epub ahead of print. PMID: 40819226. ***(Web of Science/*** *Emerging Sources Citation Index (ESCI)* ***SCOPUS-indexed)***

5. Nabilla Mohd Shukor, Al Imran Shahrul, Noraina Hafizan Norman. *Comparison of Horizontal and Vertical Tooth Movements in Erkodur vs. Zendura FLX Clear Aligners.* **Archives of Orofacial Sciences.** 2025 June: 20(1): 29-41. DOI:10.21315/aos2025.2001.OA03. ***(SCOPUS-indexed)***

6. Juliza Md Lepi, Melati Mahmud, Tong Wah Lim, Nik Mukhriz Nik Mustapha, Noraina Hafizan Norman. *Frontal Posed Smile Aesthetics in Malaysian Malay Female Patients.* EOS 2024 Abstracts, ***European Journal of Orthodontics***. 2025 April: 47(2): cjae081, DOI: 10.1093/ejo/cjae081.503 ***(Web of Science/ Science Citation Index Expanded (SCIE)/ SCOPUS-indexed)***

7. Juliza Md Lepi, Melati Mahmud, Tong Wah Lim, Nik Mukhriz Nik Mustapha, Noraina Hafizan Norman. *Photogrammetrical Investigations on Ricketts Esthetic Line In Malaysian Malay Female Adults.* EOS 2024 Abstracts, ***European Journal of Orthodontics***. 2025 April: 47(2): cjae081, DOI: 10.1093/ejo/cjae081.504 ***(Web of Science/ Science Citation Index Expanded (SCIE)/ SCOPUS-indexed)***

8. Noraina Hafizan Norman, Nurul Fazleen Che Aziz, Safiyyah Yunos, Mohd Yusmiaidil Putera Mohd Yusof, Aspalilah Alias. *Comparison of Bimaxillary Protrusion Occlusion Morphology: A Geometric Morphometric Analysis.* EOS 2024 Abstracts, ***European Journal of Orthodontics***. 2025 April: 47(2): cjae081, DOI: 10.1093/ejo/cjae081.537 ***(Web of Science/ Science Citation Index Expanded (SCIE)/ SCOPUS-indexed)***

9. Juliza Md Lepi, Melati Mahmud, Tong Wah Lim, Nik Mukhriz Nik Mustapha, Noraina Hafizan Norman. *Zero Meridien Line Norms in Malaysian Malay Female Adults.* ***2024 Dental Tribune Newspaper Zero Meridien.*** 2024 Aug: 1(1): 9-10.

10. Al Imran Shahrul , Nabilla Mohd Shukor , Noraina Hafizan Norman. *Technique for orthodontic clinical photographs.* **Cureus.** 2024 November: 16(11): e73629. DOI:10.7759/cureus.73629 ***(Web of Science/*** *Emerging Sources Citation Index (ESCI)* ***SCOPUS-indexed)***

11. Muhammad Faiz Mohd Fauad, Ku Mastura Ku Mohd Noor, Aspalilah Alias, Ker Woon Choy, Wei Lin Ng, Eric Chung, Yuan Seng Wu, Noraina Hafizan Norman. *Evaluation of Age Variation Changes in Cervical Vertebrae: 2-Dimensional (2D) Geometric Morphometrics Approach.* **Translational Research in Anatomy**. 2023 October: 33: 100269 DOI: 10.1016/j.tria.2023.100269 (***SCOPUS-indexed)***

12. Juliza Md Lepi, Melati Mahmud, Tong Wah Lim, Noraina Hafizan Norman. *Photogrammetric and Cephalometric Analyses of Ricketts' Esthetic Line in Malaysian Malay Adults: A Cross-sectional Study.* **The Open Dentistry Journal**. 2023: 17: e18742106268751 DOI: 10.2174/0118742106268751231010073305 (***SCOPUS-indexed)***

13. Tamana Sazgar, Nagham Mohammad Al-Jaf, Noraina Hafizan Norman, Aspalilah Alias. *Soft-Tissue Analysis of Different Sagittal Skeletal Patterns Using the Geometric Morphometric Method.* **European Journal of Dentistry**. 2023: 17(1): 97–106. DOI: 10.1055/s-0042-1743149. ***(Web of Science/*** *Emerging Sources Citation Index (ESCI)* ***SCOPUS-indexed)***

APPENDIX B

List of Grants

1. ***Prototype Development Research Grant Scheme (PRGS)***

   Reference: PRGS/1/2023/SKK07/UITM/02/1 (01/10/2023-30/09/2025)

   Title: Development of A Fully Automated Classification System of Cervical Vertebral Maturation Staging (CVMS) Using Deep Convolutional Neural Network.

   Members: Marshima Mohd Rosli, Mohd Yusmiaidil Putera Mohd Yusof, Noraina Hafizan Norman, Nagham Mohammed Al-Jaf, Norhasmira Mohammad

   Amount: RM210,100.00

APPENDIX C

List of Conferences

1. Oral Presentation. Title: Automated Cervical Vertebral Maturation Classification Using an Object Detection-driven Deep Convolutional Neural Network for Pubertal Growth Assessment. ***100th Annual Congress of the European Orthodontic Society, Krakow, Poland.*** 2-6th June 2025.

2. Poster Presentation. Title: Improving Cervical Vertebral Maturation Classification: A Deep Convolutional Neural Network Approach with Random Oversampling and Memory-Efficient Training. ***100th Annual Congress of the European Orthodontic Society, Krakow, Poland.*** 2-6th June 2025.

3. Participant of **2025 Three Minute Thesis Competition (3MT) Competition.** Title: Perfect Timing, Perfect Smiles: AI-Powered Growth Detection. **Shah Alam, Universiti Teknologi MARA (UiTM).** 15th May 2025.

4. Poster Presentation. Title: A Bibliometric Analysis of Artificial Intelligence in Orthodontics: Current Status, Development, and Future Research Directions. ***99th European Orthodontic Society Congress, Athens, Greece.*** 9-13th June 2024.

5. Poster Presentation. Title: Comparison of Bimaxillary Protrusion Occlusion Morphology: A Geometric Morphometric Analysis. ***99th European Orthodontic Society Congress, Athens, Greece.*** 9-13th June 2024.

6. Noraina Hafizan Norman, Marshima Mohd Rosli, Nagham Mohammed Al-Jaf, Norhasmira Mohammad, Mohd Yusmiaidil Putera Mohd Yusof. Title: Advancing Skeletal Maturity Evaluation: Utilizing Deep Convolutional Neural Networks for Cervical Vertebrae Maturation Staging ***FDCU International Symposium, Bangkok, Thailand***. 20-21st May 2024.

7. Participant of **3 Minute Thesis Competition (3MT).** Title: Predicting Cervical Bone Growth with AI: Stunted or Advanced? **Faculty of Dentistry, University Teknologi MARA (UiTM).** 21st May 2024.

8. Participant of **3 Minute Thesis Competition (3MT).** Title: Unlocking Growth Secrets: AI Predicts Bone Age with Precision. **Faculty of Dentistry, University Teknologi MARA (UiTM).** 28th February 2024.

9. Noraina Hafizan Norman, Juliza Md Lepi. Title: Photogrammetric Analyses of Zero Meridien Line in Malaysian Malay Female Adults; A Cross-Sectional Study. ***56th Annual Scientific Congress of the Korean Association of Orthodontists, Jeju, Korea***. 25-27th October 2023.

10. Poster Presentation. Title: Assessing the accuracy and reliability of the freely available software Meshmixer in measuring orthodontic dental study models. ***98th Annual Congress of the European Orthodontic Society, Oslo, Norway.*** 11-15th June 2023.

11. Poster Presentation. Title: The Cost Analysis of Using Two Clear Aligner Materials: Randomized Clinical Trial. ***98th Annual Congress of the European Orthodontic Society, Oslo, Norway.*** 11-15th June 2023.

APPENDIX D

List of Innovations

1. Noraina Hafizan Norman, Norhasmira Mohammad, Nagham Mohammed Al-Jaf, Marshima Mohd Rosli, Mohd Yusmiaidil Putera Mohd Yusof. ***Innovation title: CVMaxx: AI-Powered Growth Detection System. Invention, Innovation & Design Exposition (IIDEX 2025).*** 28-30th October 2025

2. Noraina Hafizan Norman, Norhasmira Mohammad, Nagham Mohammed Al-Jaf, Marshima Mohd Rosli, Mohd Yusmiaidil Putera Mohd Yusof. ***Innovation title: CVMaxx: AI-Powered Growth Detection System. Sarawak Digital Economy Awards.*** 4th October 2025

3. Noraina Hafizan Norman, Norhasmira Mohammad, Nagham Mohammed Al-Jaf, Marshima Mohd Rosli, Mohd Yusmiaidil Putera Mohd Yusof. ***Innovation title: CVMaxx: AI-Powered Growth Detection System. Borneo InTex International Technology Expo.*** 27-28th August 2025

4. Noraina Hafizan Norman, Norhasmira Mohammad, Nagham Mohammed Al-Jaf, Marshima Mohd Rosli, Mohd Yusmiaidil Putera Mohd Yusof. ***Awareness and Understanding of Medical Devices Quality Management System Training ISO 13485:2016.*** 1st August 2025

5. Noraina Hafizan Norman, Norhasmira Mohammad, Nagham Mohammed Al-Jaf, Marshima Mohd Rosli, Mohd Yusmiaidil Putera Mohd Yusof. ***Innovation title: CVMaxx: AI-Powered Growth Detection System. Research, Innovation and Entrepereneurship Symposium 3.0, Universiti Kebangsaan Malaysia.*** 7th May 2025

6. Noraina Hafizan Norman, Norhasmira Mohammad, Nagham Mohammed Al-Jaf, Marshima Mohd Rosli, Mohd Yusmiaidil Putera Mohd Yusof. ***Innovation title: CVMaxx: AI-Powered Growth Detection System. 50th International Exhibition of Inventions of Geneva.*** 9-13th April 2025

APPENDIX E

List of Intellectual Property Rights

1. **Copyright:**

   Notification number: **CRLY2024W09220**

   Title: CVMaXX Deep Convolutional Neural Network (CVMaXX DCNN)

   Date of Creation: 30th December 2024

2. **Trademark:**

   Trademark number: **TM2024041477 (Class 9)**

   Title: CVMaXX

   Date of Creation: 30th December 2024

3. **Patent: Utility Innovation**

   Application number: **UI2025004228**

   Title: A Method for Determining Cervical Vertebral Maturation Stages

   Date of Creation: 14th July 2025

COPYRIGHT ACT 1987

COPYRIGHT (VOLUNTARY NOTIFICATION) REGULATIONS 2012

CERTIFICATE OF COPYRIGHT NOTIFICATION
[Subregulation 8(2)]

Notification Number	:	**CRLY2024W09220**
Title of Work	:	**CVMAXX DEEP CONVOLUTIONAL NEURAL NETWORK (CVMAXX DCNN)**
Category of Work	:	**LITERARY**
Date of Notification	:	**30 DECEMBER 2024**
Date of Creation	:	**30 DECEMBER 2024**

This is to certify, under the Copyright Act 1987 [Act 332] and the Copyright (Voluntary Notification) Regulations 2012 that the copyrighted work bearing the Notification No. above for the applicant **UNIVERSITI TEKNOLOGI MARA** as the **OWNER** and **MOHD YUSMIAIDIL PUTERA MOHD YUSOF (**[redacted]**), MARSHIMA MOHD ROSLI (**[redacted]**), NORAINA HAFIZAN BINTI NORMAN (**[redacted]**), NAGHAM MOHAMMED AL-JAF (**[redacted]**), NORHASMIRA MOHAMMAD (**[redacted]**)** as the **AUTHOR** have been recorded in the Register of Copyright, in accordance with section 26B of the Copyright Act 1987 [Act 332].

KAMAL BIN KORMIN
CONTROLLER OF COPYRIGHT
MALAYSIA

(Agency under the Ministry of Domestic Trade and Cost of Living)

PERBADANAN HARTA INTELEK MALAYSIA
INTELLECTUAL PROPERTY CORPORATION OF MALAYSIA
(Agensi dibawah KPDN)

Aras LG, G, 2-5, 11-13 & 15-23,
Menara MyIPO, PJ Sentral,
Lot 12, Persiaran Barat, Seksyen 52,
46200 Petaling Jaya, Selangor,
MALAYSIA

Tel : +603 – 7496 8900
Faks(Fax) : +603 – 7496 8999
Laman Web (Web) : www.myipo.gov.my

NOTIS SETUJUTERIMA

Rujukan Tuan :
Tarikh : **03 September 2025**

UNIVERSITI TEKNOLOGI MARA
Faculty of Dentistry Universiti Teknologi Mara,
Jalan Hospital, Kampus Sg. Buloh,
47000 Sungai Buloh Selangor,
Malaysia.

NO. PERMOHONAN : **TM2024041477**
KELAS : **9**
PEMOHON/PEMILIK : **UNIVERSITI TEKNOLOGI MARA**
Faculty of Dentistry Universiti Teknologi Mara,
Jalan Hospital, Kampus Sg. Buloh,
47000 Sungai Buloh Selangor,
Malaysia.

Dengan hormatnya saya diarah merujuk kepada permohonan Cap Dagangan di atas.

2. Semakan ke atas permohonan tuan telah dibuat dan **disetujuterima**. Permohonan ini akan disiarkan dalam tempoh masa yang terdekat bagi tujuan bangkangan. Semakan penyiaran boleh dibuat di *ipjournal.myipo.gov.my* pada setiap hari **Khamis**.

3. Sebarang pertanyaan berkaitan penyiaran Cap Dagangan boleh menghubungi Puan Zakini binti Yaman di talian 03-74968952 atau e-mel zakini@myipo.gov.my.

4. Notis pendaftaran Cap Dagangan akan dikeluarkan selepas 2 bulan 2 minggu daripada tarikh siaran sekiranya permohonan ini tiada bangkangan oleh mana-mana pihak berkepentingan. Tuan boleh failkan borang TMJ4 (fi: RM50.00) untuk mendapat suatu perakuan pendaftaran (sijil). Pertanyaan berkaitan notis pendaftaran boleh menghubungi Puan Zaiton Binti Haris di talian 03-74968941 atau emel zaitonh@myipo.gov.my

Sekian, terima kasih.

"MALAYSIA MADANI"

"BERKHIDMAT UNTUK NEGARA"

Saya yang menjalankan amanah,

(MOHAMMAD FATHIR BIN MD NOOR)
b/p Pendaftar Cap Dagangan Malaysia
T:03-7496 3843 | E: fathir@myipo.gov.my

NOTIFICATION OF THE REGISTRATION OF TRADEMARK

Trademark No.	:	**TM2024041477**
Registered Proprietor	:	**UNIVERSITI TEKNOLOGI MARA**
Date of Registration	:	**30 DECEMBER 2024**
Renewal Due Date	:	**30 DECEMBER 2034**

CVMaXX

The above-numbered trademark has been registered and the above-named has been entered as the proprietor of the said trademark in the Register in respect of the following goods or services [with condition(s)]:

(LIST OF GOOD/SERVICES : AS ATTACHED)

Registrar of Trademarks

PERBADANAN HARTA INTELEK MALAYSIA
INTELLECTUAL PROPERTY CORPORATION OF MALAYSIA
(Agensi di bawah KPDN)
Aras LG, G, 2-5, 11-13 & 15-23,
Menara MyIPO, PJ Sentral,
Lot 12, Persiaran Barat, Seksyen 52,
46200 Petaling Jaya, Selangor,
MALAYSIA

Tel : +603 – 7496 8900
Faks(Fax) : +603 – 7496 8999
Laman Web (Web) : www.myipo.gov.my

CERTIFICATE OF FILING
(REGULATION 25(1))

APPLICANT	:	UNIVERSITI TEKNOLOGI MARA
APPLICATION NO	:	UI2025004228
RECEIVED ON	:	14 JULY 2025
FILING DATE	:	14 JULY 2025
AGENT'S/APPLICANT'S FILE REF.	:	2025/UI/Q6.10/OP

Please take note that the applicant is required to comply with the requirements of the Patents Act 1983 and Regulations.

Remarks:
Number of Claims Approved: 1

Date : 29 July 2025

(ASMAWATI JUSOH)
For Registrar of Patents
asmawati@myipo.gov.my
03-7496 3295

To : TEE LIN YIK
C/O TEE IP SDN BHD
A-23-01, PENTHOUSE OFFICE EKOCHERAS OFFICE SUITES
693, JALAN CHERAS, BATU 5
56000 CHERAS
KUALA LUMPUR
MALAYSIA

ADVICE TO APPLICANTS - TIME PERIODS/REQUEST FOR EXAMINATION

During the prosecution of an application, the applicant is required to take certain actions within specified time period. If no action is taken, the application may be refused or considered withdrawn. The attention of applicant is drawn, to the following:

(a) A request for Substantive Examination should on Form 5, together with prescribed fee, **within 18 months from the filing date of the application, otherwise the application may be treated as withdrawn.**

OR

(b) A request for Modified Substantive Examination should be made on Form 5A, together with prescribed fee, **within 18 months from the filing date of the application, otherwise the application may be treated as withdrawn.**

In accordance with Section 29A(8) of the Act the 18 months period cannot be extended under the provisions of Section 82. However, deferment of Modified Substantive Examination may be requested under Section 29A(6) and the request for deferment must be filed within the 18 months period.

For divisional application: -

A request for Substantive Examination shall be made on Form 5, together with prescribed fee, at the time of the filing of the request for the division of application, **otherwise the application may be treated as withdrawn.**

OR

A request for Modified Examination shall be made on Form 5A, together with prescribed fee, at the time of the filing of the request for the division of application, **otherwise the application may be treated as withdrawn**

APPENDIX E

List of Awards

1. ***Gold Award*** for invention ***'CVMaxx: AI-Powered Growth Detection System'*** at the Invention, Innovation & Design Exposition (IIDEX 2025), Universiti Teknologi MARA, Shah Alam, Malaysia. 28-30th October 2025

2. ***Gold Award*** for invention ***'CVMaxx: AI-Powered Growth Detection System'*** at the Borneo InTex International Technology Expo, Sarawak, Malaysia. 27-28th August 2025

3. ***Gold Award with Jury Recommendations*** for invention ***'CVMaxx: AI-Powered Growth Detection System'*** at the 50th International Exhibition of Inventions of Geneva, Geneva, Switzerland. 9-13th April 2025

4. ***Thailand Award for the Best International Invention and Innovation*** *for invention* ***'CVMAXX'*** by the National Research Council of Thailand at 50th International Exhibition of Inventions of Geneva, Geneva, Switzerland. 9-13th April 2025

5. ***Distinguished Innovation Award*** *by King Abdulaziz University for invention* ***'CVMAXX'*** at the 50th International Exhibition of Inventions of Geneva, Geneva, Switzerland. 9-13th April 2025

6. ***Second Prize Award*** for 3MT PhD Category presentation entitled ***'Unlocking Growth Secrets: AI Predicts Bone Age with Precision'*** at the 1st UiTM Dental Postgraduate Scientific Symposium 2024 Universiti Teknologi MARA, Sg Buloh, Malaysia. 25th May 2024.

7. ***Bronze Award*** for invention ***'AIDentify: Enhancing Dental Care with an AI-Driven Colour Coded System'*** at the International Invention & Innovation in Dentistry Exhibition (IIIDentEx 2024)' Universiti Teknologi MARA, Sungai Buloh Campus, Malaysia. 19th October 2024.


iidex 2025
Invention, Innovation & Design Exposition
This certificate is presented to
MOHD YUSMIAIDIL PUTERA B. MOHD YUSOF
NORAINA HAFIZAN NORMAN
MARSHIMA MOHD ROSLI
NAGHAM MOHAMMED AL-JAF
NORHASMIRA MOHAMMAD
has been awarded
GOLD AWARD
for the Invention/Innovation/Design of
Title: CVMaXX – AI-Powered Growth Detection System
at
INVENTION, INNOVATION & DESIGN EXPOSITION (iidex) 2025
28- 30 October 2025
Dewan Agung Tuanku Canselor (DATC)
UiTM Shah Alam, Selangor, Malaysia
PROF. TS. DR. NORAZAH ABD RAHMAN
Deputy Vice-Chancellor (Research & Innovation)
Universiti Teknologi MARA
Organised by
Department of Research & Innovation, UiTM
Certificate of
GOLD AWARD


Borneo InTEX
INTERNATIONAL TECHNOLOGY EXPO
Bridging Ideas: Research, Industry and Community in Action
27th - 28th August 2025
Raia Hotel Convention Centre Kuching, SARAWAK
Gold Award
AWARDED TO
Norhasmira binti Mohammad
Mohd Yusmiadil Putera Mohd Yusof, Nagham Mohammed Abdullah, Marshima Binti Mohd Rosli, Noraina Hafizan Norman
FOR OUTSTANDING INVENTION / INNOVATION OF
CVMaXX©: AI-Powered Growth Detection System
Prof. Dr Awang Ahmad Sallehin bin Awang Husaini
Chairperson for Borneo InTEX2025
Centre For Research, Development, Innovation, Commercialisation & Economy
Organised by
UNIMAS
Supported by
SARAWAK
UNIMAS
UNIVERSITI MALAYSIA SARAWAK
UNIMAS INNOVATION
Strategic Partners

DIPLÔME

inventions Geneva

SALON INTERNATIONAL DES INVENTIONS GENÈVE

Après examen, le Jury International a décidé

de remettre à : Noraina NORMAN, Norhasmira MOHAMMAD, M. Yusmiaidil P. M. YUSOF, Marshima M. ROSLI, Nagham M. AL-JAF, UITM

pour l'invention : CVMaXX : système avancé de classification basé sur les données pour le stade de maturation vertébrale cervicale

MÉDAILLE D'OR
GOLD MEDAL
GOLDMEDAILLE

Avec les félicitations du jury
With the congratulations of the jury
Mit höchsten Empfehlungen des Preisgerichtes

Genève, le 12 avril 2025

Le Président du Jury: David Taji

Le Président du Salon: Jean-Luc Vincent

National Research Council of Thailand

THAILAND AWARD FOR THE BEST INTERNATIONAL INVENTION & INNOVATION

Presented to

Noraina Norman, Mohd Yusmiaidil Putera M. Yusof, Norhasmira Mohammad

For the excellent invention

CVMAXX

In "The 50th International Exhibition of Inventions Geneva"
Geneva, Switzerland
9 - 13 April 2025

(Dr. Wiparat De-ong)
Executive Director
National Research Council of Thailand

King Abdulaziz University

Cordially Presents this

Distinguished Innovation Award

to

CVMAXX

by

NORAINA NORMAN, MOHD YUSMIAIDIL PUTERA M. YUSOFF ,NORHASMIRA MOHAMAD

from

Universiti Teknologi MARA

Exhibited at

50th International Exhibition of Inventions of Geneva
April 9 – 13, 2025

Hossam Khalifa
Talent and Creative Center


UNIVERSITI TEKNOLOGI MARA
Fakulti Pergigian
25 TAHUN
1999-2024
UiTM SEBUAH UNIVERSITI
CERTIFICATE
OF ACHIEVEMENT
proudly presented to
AP NORAINA HAFIZAN NORMAN
for being 2nd place 3MT (PHD) in the
1ST UiTM DENTAL POSTGRADUATE SCIENTIFIC SYMPOSIUM (DPgSS)
16-17 JULY 2024
Prof Dr Nur Aida Asyikin Abd Rahman
Dean
Faculty of Dentistry
Universiti Teknologi Mara
Dr Marlena Kamaruzaman
Program Director
1st UiTM Dental Postgraduate Scientific Symposium (DPgSS)

CERTIFICATE
Of Achievement

This certifies that

TEAM MEMBERS: **Asruni Ab Ghani, Hafizul Izwan Mohd Zahari, Noraina Hafizan Norman, Nur Hafizah Kamar Affendi, Melati Mahmud**

Has been awarded

BRONZE AWARD

For the project entitled

AIDentify: Enhancing Dental Care with an AI-Driven Color Coded System

At the

International Invention & Innovation in Dentistry Exhibition 2024

Held at Noble Resort Hotel, Melaka
On 19th October, 2024

Professor Dr. Aida Nur Ashikin Abd Rahman
DEAN
FACULTY OF DENTISTRY
UNIVERSITI TEKNOLOGI MARA

IIIDentEx2024

HSDC RESOURCES (MA0207073-X)

Certificate

THIS IS TO CERTIFY THAT

DR. NORAINA HAFIZAN NORMAN

HAS SUCCESSFULLY COMPLETED THE TRAINING COURSE FOR

AWARENESS AND UNDERSTANDING OF MEDICAL DEVICES QUALITY MGMT. SYSTEM ISO 13485:2016

COURSE DATE - 01st August 2025

CERTIFICATE NO. - HSDC/PCSB/MDQMS/08/25-001

ORGANISATION - PRESIDENTAL CARE SDN. BHD.

HSDC RESOURCES

COURSE FACILITATOR

Delegasi Malaysia menang dua emas, tiga perak dan satu gangsa di Geneva

April 15, 2025 @ 2:56pm

PEMENANG pingat emas yang membangunkan produk inovasi CVMAXX dari UiTM, Dr Noraina Norman dan Dr Norhasmira Mohamad.

Kuala Lumpur: Delegasi Malaysia meraih dua pingat emas, tiga perak dan satu gangsa di Pameran Rekacipta dan Inovasi Antarabangsa ke-50 di Geneva, Switzerland berlangsung 9 hingga 13 April lalu.

Pingat emas dimenangi pereka cipta dari Universiti Teknologi MARA (UiTM) iaitu Dr Noraina Norman, Dr Norhasmira Mohamad dan Prof Dr Muhamad Yusmiaidil Putera M Yusof yang membangunkan produk inovasi diberi nama CVMAXX.

Menurut mereka, CVMAXX adalah sistem pintar berasaskan kecerdasan buatan (AI) yang membantu menilai tahap kematangan vertebra servikal secara automatik.

"Sistem ini mengurangkan ralat manusia, meningkatkan ketepatan penilaian, mengoptimumkan penentuan masa rawatan serta menyumbang kepada penambah baikan hasil rawatan dalam bidang kesihatan," katanya.

Satu lagi pingat emas dimenangi produk inovasi diberi nama Kiker's Stick yang dihasilkan oleh Nik Ahmad Faris Nooh dan rakan dari Jabatan Kebajikan Masyarakat (JKM). Menurutnya, Kiker's Stick adalah alat bantuan untuk pesakit angin ahmar atau warga emas yang mempunyai masalah kestabilan dan keseimbangan badan ketika berjalan.

"Tayar dan brek yang terdapat di Kiker's Stick membantu pesakit lumpuh separuh badan (angin ahmar) untuk berjalan. Mereka tidak perlu mengangkat tongkat ketika berjalan dan mempunyai ciri keselamatan iaitu brek ketika mahu berhenti bagi mengelak pesakit jatuh. "Alat ini juga membantu untuk menguatkan otot, sendi kaki dan meningkatkan keseimbangan badan ketika berjalan. Ia juga dapat menjimatkan masa pengguna untuk bergerak dari satu tempat ke satu tempat," katanya.

DELEGASI Malaysia di Pameran Rekacipta dan Inovasi Antarabangsa ke-50 di Geneva, Switzerland.

Produk reka cipta yang memenangi hadiah di peringkat antarabangsa ini mempunyai ketulenan tersendiri, unik, bernilai komersial dan membantu dalam kelestarian hidup masyarakat. Pameran Rekacipta dan Inovasi Antarabangsa kali ini disertai 52 negara dan menerima 1,050 penyertaan reka cipta.

Selain Malaysia, turut mengambil bahagian ialah China, Arab Saudi, Thailand, Jepun, Taiwan, Korea Selatan dan lain-lain. Selain UiTM dan JKM, Malaysia turut diwakili pereka cipta dari Majlis Bandaraya Kuantan, DINA Resources & Management Sdn Bhd dan Adastra IP Sdn Bhd. Penyertaan delegasi Malaysia ke acara itu diurus dan dibimbing oleh Wan Global Invention & Innovation Enterprise yang diketuai Dr Wan Manshol W Zin.

AUTHOR'S PROFILE

Noraina Hafizan Norman is a specialist in Orthodontics at the Faculty of Dentistry, Universiti Teknologi MARA (UiTM), where she is actively involved in clinical practice, teaching, and postgraduate supervision. She completed her specialty training in Orthodontics at the University of Manchester Dental Hospital and holds Membership in Orthodontics (MOrth) from both the Royal College of Surgeons of England and the Royal College of Physicians and Surgeons of Glasgow.

Her academic and clinical interests include cephalometric analysis, orthodontic treatment outcomes, and imaging-based assessments. Her recent research explores the integration of artificial intelligence in orthodontic diagnostics and cervical vertebral maturation assessment. She has published in peer-reviewed journals, supervised postgraduate research, and contributed to professional and academic orthodontic activities at both national and international levels.

*ATTACHMENT:

TITLE: DEVELOPMENT AND VALIDATION OF AN AUTOMATED DEEP LEARNING-BASED CERVICAL VERTEBRAL MATURATION CLASSIFICATION SYSTEM USING OBJECT DETECTION AND MORPHOLOGICAL ANALYSIS

MAIN SUPERVISOR / CAMPUS: PROF DR MOHD YUSMIAIDIL PUTERA MOHD YUSOF / SUNGAI BULOH CAMPUS