



A Prospective Multicenter Simulation Study

Artificial Intelligence–Assisted Clinical Decision Support for Personalized Bimaxillary Orthognathic Surgery

Ayhan Yildirim^{1*}, René Hertach², Vedat Yildirim³

¹ Seeklinik Zurich, Department of Oral and Maxillofacial Surgery and Clinical Research, Zurich, Switzerland

² Center for Prosthodontics and Digital Dentistry, Zurich, Switzerland

³ Munich Center for Oral and Maxillofacial Surgery, Munich, Germany

ORCID IDs

● Ayhan Yildirim: 0009-0009-2179-1802

● Vedat Yildirim: 0009-0009-8664-5953

* Correspondence to: Prof. Dr. Dr. Yildirim, Hochschule Zürich – Independent Academy for Medicine and Dentistry
Albisstrasse 80, 8038 Zurich, Switzerland, E-mail: ayhan.yildirim@hs-zh.ch

ARTICLE INFO

Article history:

Received: 12 March 2026, Revised: 24 April 2026, Accepted: 05 May 2026, Available online: 05 June 2026,
Version of Record: 05 June 2026.

© The Author(s) 2026. Published by Journal of Medicine and Dentistry (JMDNT).

This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0).

ABSTRACT

Background

Artificial intelligence (AI) has rapidly evolved from automated image analysis toward advanced clinical decision-support systems capable of assisting surgeons throughout the treatment planning process. Although recent studies have demonstrated high accuracy of AI-assisted virtual surgical planning and postoperative outcome prediction in orthognathic surgery, its ability to support individualized surgical decision-making has

not been systematically investigated. The aim of the present study was to evaluate the performance of a multimodal AI-assisted clinical decision-support system for personalized treatment planning in patients undergoing bimaxillary orthognathic surgery.

Methods

A prospective multicenter simulation study was conducted using standardized digital datasets from 148 consecutive patients who underwent bimaxillary orthognathic surgery at two European oral and maxillofacial surgery centers. The AI platform integrated cone-beam computed tomography, intraoral digital scans, automated cephalometric analysis, three-dimensional facial morphology, demographic characteristics, and virtual surgical planning data into a multimodal deep learning framework. The clinical decision-support system automatically generated individualized surgical recommendations regarding maxillary and mandibular repositioning, rotational correction, genioplasty, and predicted postoperative facial aesthetics, skeletal stability, and relapse risk. AI-generated treatment strategies were compared with consensus recommendations from an expert panel consisting of experienced oral and maxillofacial surgeons.

Results

The AI-assisted clinical decision-support system demonstrated excellent agreement with expert treatment recommendations, achieving an overall decision concordance of 94.6%. Mean differences between AI-recommended and expert-planned skeletal movements measured 0.8 ± 0.4 mm for translational movements and $0.6 \pm 0.3^\circ$ for rotational corrections. AI-assisted treatment planning reduced mean planning time by 58% while maintaining high reproducibility across repeated analyses (ICC = 0.991). The integrated prediction framework estimated postoperative facial morphology with a mean surface deviation of 0.93 ± 0.35 mm and predicted skeletal relapse with an area under the receiver operating characteristic curve of 0.953. Explainable AI analysis demonstrated that skeletal discrepancy, facial asymmetry, cephalometric relationships, and planned mandibular advancement represented the most influential variables guiding individualized treatment recommendations.

Conclusions

Multimodal artificial intelligence demonstrated excellent performance as a clinical decision-support system for personalized bimaxillary orthognathic surgery. By integrating automated diagnosis, individualized surgical planning, postoperative outcome prediction, and explainable decision-making into a unified workflow, AI has the potential to enhance precision surgery, improve treatment standardization, and support multidisciplinary clinical decision-making. Prospective clinical validation is warranted before routine implementation in daily practice.

Keywords

Artificial Intelligence; Clinical Decision Support; Orthognathic Surgery; Bimaxillary Osteotomy; Personalized Surgery; Deep Learning; Virtual Surgical Planning; Explainable Artificial Intelligence; Precision Medicine; Maxillofacial Surgery.

1. INTRODUCTION

Orthognathic surgery represents the standard treatment for patients with moderate to severe dentofacial deformities that cannot be adequately corrected by orthodontic treatment alone. During the past three decades, substantial advances in surgical techniques, fixation systems, perioperative orthodontics, and digital technologies have markedly improved the predictability, safety, and functional outcomes of bimaxillary surgery. Contemporary treatment objectives extend beyond the correction of skeletal malocclusion and occlusal discrepancies and increasingly emphasize facial aesthetics, functional rehabilitation, airway optimization, long-term skeletal stability, and improvement of oral health-related quality of life [1–6].

The introduction of cone-beam computed tomography (CBCT), three-dimensional (3D) facial imaging, intraoral optical scanning, computer-aided design and manufacturing (CAD/CAM), and virtual surgical planning (VSP) has fundamentally transformed orthognathic surgery. Compared with conventional model surgery, digital workflows allow precise visualization of craniofacial anatomy, simulation of osteotomies,

fabrication of patient-specific splints and guides, and accurate transfer of virtual treatment plans to the operating room. Numerous clinical studies have demonstrated that VSP improves surgical precision while reducing planning variability and enhancing interdisciplinary communication between surgeons, orthodontists, and biomedical engineers [1–13].

Despite these technological advances, treatment planning remains a highly complex cognitive process that depends largely on surgeon experience. Establishing an optimal surgical strategy requires simultaneous consideration of multiple anatomical, functional, biomechanical, and aesthetic variables. The surgeon must determine the extent of maxillary advancement or impaction, the magnitude of mandibular repositioning, rotational correction of the maxillomandibular complex, the indication for adjunctive genioplasty, and the anticipated influence of each movement on facial appearance, skeletal stability, occlusion, airway dimensions, and long-term treatment outcomes. Because these variables interact in a highly nonlinear manner, treatment planning frequently involves iterative discussions among surgeons, orthodontists, and laboratory engineers before a final surgical strategy is selected [2–10].

Although virtual surgical planning provides an accurate geometric representation of craniofacial anatomy, current software systems primarily function as simulation platforms rather than true clinical decision-support systems. Existing planning software allows surgeons to manually modify skeletal movements and visualize the resulting anatomical changes; however, the software does not objectively evaluate alternative treatment strategies or recommend the most appropriate surgical approach for an individual patient. Consequently, the final treatment decision continues to rely almost exclusively on clinical experience and subjective interpretation of multiple imaging datasets.

This limitation has become increasingly relevant as treatment objectives have shifted toward personalized medicine. Patients presenting with similar cephalometric measurements may require substantially different surgical strategies depending on facial proportions, soft tissue morphology, skeletal asymmetry, temporomandibular joint status, airway characteristics, age, sex, and aesthetic expectations. Consequently, individualized treatment planning has emerged as one of the principal challenges of contemporary orthognathic surgery.

Artificial intelligence (AI) has recently emerged as one of the most rapidly developing technologies in medicine and has demonstrated remarkable potential for improving diagnostic accuracy, workflow automation, predictive analytics, and clinical decision-making. In contrast to conventional rule-based software, modern deep learning algorithms are capable of identifying highly complex nonlinear relationships within large multimodal datasets, enabling individualized prediction of clinical outcomes that cannot easily be achieved using traditional statistical approaches [14–18]. Within radiology, pathology, oncology, cardiology, and orthopedic surgery, AI-assisted systems have already demonstrated diagnostic performance comparable to experienced specialists while substantially reducing analysis time and improving reproducibility [14–18].

Within oral and maxillofacial surgery, the application of artificial intelligence has evolved rapidly over the last few years. Initial investigations focused primarily on automated detection of mandibular and midfacial fractures, segmentation of craniofacial structures, and automated interpretation of CBCT datasets [19–26]. More recently, AI-assisted workflows have been successfully applied to virtual surgical planning, automated cephalometric landmark detection, postoperative soft tissue prediction, prediction of skeletal relapse, and multicenter validation of digital orthognathic surgery workflows [27–33]. Collectively, these investigations demonstrated that artificial intelligence can substantially improve efficiency, reproducibility, and predictive accuracy throughout the digital treatment pathway.

Nevertheless, most currently available AI applications remain task-specific. Existing algorithms generally perform isolated functions such as image segmentation, cephalometric landmark identification, postoperative outcome prediction, or automated measurements. Although these individual components contribute to workflow optimization, they do not actively support the surgeon during the most critical phase of treatment planning—the selection of the optimal surgical strategy. Consequently, a substantial gap remains between automated image analysis and true AI-assisted clinical decision-making.

Clinical decision-support systems (CDSS) represent the next evolutionary step in medical artificial intelligence. Rather than replacing clinical judgment, CDSS integrate heterogeneous sources of patient-specific information to generate evidence-based recommendations that assist clinicians during complex therapeutic decisions. Such systems have already demonstrated clinical benefit in intensive care medicine, oncology, cardiovascular medicine, and radiology by improving treatment standardization, reducing diagnostic variability, and facilitating individualized patient management [14–18]. However, comparable decision-support frameworks have not yet been systematically investigated for orthognathic surgery.

The rapid evolution of artificial intelligence has fundamentally changed the role of computational technologies in medicine. Whereas early machine learning applications primarily focused on image classification and automated detection tasks, contemporary AI systems increasingly support complex clinical reasoning by integrating heterogeneous patient information into individualized decision-making frameworks. Recent advances in multimodal deep learning, transformer-based architectures, and foundation models have enabled simultaneous analysis of imaging data, clinical variables, laboratory findings, and patient-specific characteristics, thereby extending artificial intelligence beyond diagnosis toward comprehensive clinical decision support [14–18,34–38].

In orthognathic surgery, such a transition is particularly relevant because treatment planning represents a multidimensional optimization problem rather than a purely anatomical correction. The surgeon must continuously balance competing objectives, including postoperative facial aesthetics, functional occlusion, airway dimensions, skeletal stability, temporomandibular joint health, soft tissue adaptation, and long-term patient satisfaction. Optimization of one parameter may negatively influence another, making treatment planning highly dependent on clinical experience and multidisciplinary discussion. Consequently, individualized treatment planning cannot be reduced to isolated cephalometric measurements or predefined treatment algorithms but requires simultaneous integration of numerous anatomical, biomechanical, and patient-specific variables.

Recent investigations from our research group have demonstrated that artificial intelligence can reliably automate many individual components of this workflow. Deep learning algorithms successfully performed automated craniofacial segmentation, cephalometric landmark detection, virtual surgical planning, postoperative soft tissue prediction, skeletal relapse prediction, multicenter workflow validation, and health economic optimization of AI-assisted orthognathic surgery [19–33]. Collectively, these studies established the technical feasibility and clinical robustness of AI-assisted digital workflows while demonstrating substantial improvements in efficiency, reproducibility, and predictive accuracy. However, although these investigations substantially advanced digital treatment planning, they intentionally focused on individual workflow components rather than integrated clinical decision-making.

The next logical step therefore involves transforming artificial intelligence from an analytical tool into an intelligent clinical assistant capable of supporting therapeutic decision-making. Clinical decision-support systems (CDSS) have already demonstrated considerable benefit across multiple medical specialties. In oncology, AI-assisted decision-support platforms integrate imaging findings, histopathology, molecular biomarkers, and clinical guidelines to recommend individualized treatment strategies. Comparable systems have improved antibiotic stewardship in infectious diseases, cardiovascular risk stratification, intensive care management, and radiological reporting by reducing interobserver variability and increasing adherence to evidence-based recommendations [14–18,39]. Despite these advances, equivalent decision-support systems have not yet been systematically investigated for orthognathic surgery.

An effective AI-assisted decision-support system should not merely recommend surgical movements but should evaluate alternative treatment strategies according to multiple clinically relevant outcome measures. Such a system may simultaneously estimate postoperative facial aesthetics, skeletal stability, relapse probability, airway changes, functional occlusion, facial symmetry, and anticipated patient satisfaction before recommending an individualized surgical plan. Rather than replacing surgeon expertise, artificial intelligence may function as a collaborative decision-support partner that objectively evaluates complex treatment scenarios while leaving the final therapeutic decision to the treating surgeon. This concept of human–AI collaboration is increasingly regarded as the most realistic and clinically acceptable model for implementation of artificial intelligence within healthcare.

Another essential prerequisite for successful clinical implementation is explainability. Although deep neural networks frequently achieve outstanding predictive performance, their clinical acceptance remains limited when treatment recommendations cannot be adequately explained. The emergence of Explainable Artificial Intelligence (XAI) has therefore become one of the most important developments in contemporary medical AI research. Modern explainability techniques—including attention mechanisms, SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (Grad-CAM)—allow visualization of the variables that contribute most strongly to AI-generated recommendations. Such approaches improve clinician confidence, facilitate regulatory approval, and enhance transparency of AI-assisted decision-making [40–43]. Within orthognathic surgery, explainable AI may allow surgeons to understand why a particular patient is recommended to undergo genioplasty, counterclockwise rotation of the maxillomandibular complex, or alternative skeletal movements, thereby transforming AI from an opaque prediction model into an interpretable clinical assistant.

These developments closely align with the broader concepts of precision surgery and digital twins. Precision surgery aims to individualize operative treatment according to patient-specific anatomical, functional, and biological characteristics rather than relying on standardized treatment algorithms. Digital twins extend this concept further by creating continuously evolving virtual patient models integrating imaging data, surgical simulations, biomechanical analyses, and longitudinal clinical information into a unified computational framework. Artificial intelligence represents the central enabling technology underlying these concepts because it provides the computational capability required to continuously analyze large multimodal datasets and generate individualized treatment recommendations [44–49].

The present study builds directly upon our previous investigations in artificial intelligence-assisted orthognathic surgery while addressing an important remaining gap in the literature. Rather than evaluating isolated components of digital treatment planning, we developed a multimodal AI-assisted clinical decision-support system designed to recommend individualized surgical strategies for patients undergoing bimaxillary orthognathic surgery. The proposed framework integrates automated cephalometric analysis, three-dimensional craniofacial imaging, facial morphology, demographic variables, virtual surgical planning, postoperative outcome prediction, and explainable artificial intelligence within a unified decision-support architecture.

We hypothesized that a multimodal AI-assisted clinical decision-support system would demonstrate excellent agreement with expert treatment recommendations while simultaneously improving planning efficiency, reducing interobserver variability, and supporting personalized surgical decision-making. Furthermore, we hypothesized that explainable AI techniques would provide transparent visualization of the clinical variables underlying individualized treatment recommendations, thereby enhancing interpretability and facilitating future clinical implementation.

2. MATERIALS AND METHODS

2.1 Study Design and Ethical Considerations

The present investigation was designed as a prospective multicenter simulation study evaluating the performance of an artificial intelligence-assisted clinical decision-support system for personalized treatment planning in patients undergoing bimaxillary orthognathic surgery. The study was performed collaboratively between the Departments of Oral and Maxillofacial Surgery at two tertiary referral centers with extensive experience in digital orthognathic surgery. All digital workflows, imaging protocols, and AI analyses were standardized across both institutions before initiation of the study to ensure methodological consistency and reproducibility.

Because the present investigation was conducted as a simulation study using fully anonymized digital datasets without influencing clinical patient management, no experimental intervention was performed on patients. The simulated AI-assisted treatment recommendations were generated independently from the original clinical

treatment plans and therefore had no influence on patient care. All datasets were anonymized before analysis in accordance with institutional data protection regulations and the principles of the Declaration of Helsinki.

The study protocol followed the recommendations of the STROBE statement for observational studies and incorporated the recently published CLAIM (Checklist for Artificial Intelligence in Medical Imaging) recommendations to improve transparency and reproducibility of artificial intelligence research. Furthermore, methodological reporting was aligned with current recommendations for machine learning studies in surgery, including detailed documentation of model development, validation strategy, performance metrics, and explainability analysis.

The primary objective of the study was to evaluate the agreement between AI-generated treatment recommendations and consensus recommendations established by an independent panel of experienced oral and maxillofacial surgeons. Secondary objectives included assessment of planning efficiency, reproducibility, explainability of AI-generated recommendations, prediction of postoperative facial aesthetics and skeletal stability, and overall clinical utility of the proposed decision-support framework.

2.2 Study Population

Digital records from 148 consecutive patients who underwent primary bimaxillary orthognathic surgery between January 2021 and December 2025 were retrospectively collected from institutional databases and prospectively processed within the standardized AI workflow. Patient selection was performed consecutively to minimize selection bias. All cases represented completed orthognathic treatment including preoperative orthodontics, virtual surgical planning, bimaxillary osteotomy, and postoperative follow-up.

Eligible patients were required to be at least 18 years of age and present with skeletal Class II or skeletal Class III dentofacial deformities requiring combined Le Fort I osteotomy and bilateral sagittal split osteotomy. All patients underwent standardized preoperative CBCT imaging, intraoral digital scanning, clinical photography, and comprehensive cephalometric analysis before virtual surgical planning.

Patients presenting with syndromic craniofacial deformities, cleft lip and palate, craniofacial trauma, distraction osteogenesis, previous orthognathic surgery, temporomandibular joint replacement, incomplete digital imaging datasets, or severe CBCT artifacts impairing automated image processing were excluded from the study. Patients with missing postoperative imaging or incomplete clinical documentation were likewise excluded to ensure complete digital datasets for simulation.

The final study cohort consisted of 148 patients, including 78 skeletal Class II and 70 skeletal Class III deformities. Mean patient age was 28.7 ± 6.4 years, and the cohort comprised 94 female and 54 male patients. Baseline demographic characteristics, skeletal classification, cephalometric parameters, and surgical movement characteristics are summarized in [Table 1](#).

Table 1. Demographic and clinical characteristics of the study population (n = 148).				
Characteristic	Total (n = 148)	Skeletal Class II (n = 78)	Skeletal Class III (n = 70)	p-value
Demographic characteristics				
Age, years, mean $\pm$ SD (range)	28.7 $\pm$ 6.4 (18-52)	28.4 $\pm$ 6.1 (18-49)	29.0 $\pm$ 6.8 (18-52)	0.58
Female, n (%)	94 (63.5)	52 (66.7)	42 (60.0)	0.37
Male, n (%)	54 (36.5)	26 (33.3)	28 (40.0)	0.37
Body mass index, kg/m ² , mean $\pm$ SD	23.8 $\pm$ 3.1	23.6 $\pm$ 2.9	24.0 $\pm$ 3.3	0.45
Skeletal classification, n (%)				
Skeletal Class II	78 (52.7)	–	–	–
Skeletal Class III	70 (47.3)	–	–	–
Preoperative cephalometric values, mean $\pm$ SD				
ANB angle, °	–0.6 $\pm$ 2.6	4.3 $\pm$ 1.8	–5.7 $\pm$ 2.0	<0.001*
SNA angle, °	81.4 $\pm$ 3.2	83.4 $\pm$ 2.5	79.1 $\pm$ 2.6	<0.001*
SNB angle, °	82.0 $\pm$ 3.4	79.3 $\pm$ 2.4	85.0 $\pm$ 2.6	<0.001*
Wits appraisal, mm	–0.8 $\pm$ 2.7	2.0 $\pm$ 1.8	–4.0 $\pm$ 1.9	<0.001*
Lower facial height (ANS-Me), mm	65.2 $\pm$ 5.0	64.7 $\pm$ 4.7	65.8 $\pm$ 5.3	0.18
Mandibular plane angle (SN-GoGn), °	31.5 $\pm$ 6.0	30.4 $\pm$ 5.6	32.7 $\pm$ 6.3	0.02*
Planned skeletal movements (AI-assessed), mean $\pm$ SD				
Maxillary advancement, mm	3.9 $\pm$ 2.2	4.3 $\pm$ 2.1	3.4 $\pm$ 2.1	0.01*
Maxillary impaction, mm	1.8 $\pm$ 1.6	1.4 $\pm$ 1.4	2.2 $\pm$ 1.7	0.002*
Mandibular advancement, mm	5.1 $\pm$ 2.6	5.9 $\pm$ 2.2	4.2 $\pm$ 2.4	<0.001*
Mandibular setback, mm	1.6 $\pm$ 1.8	0.4 $\pm$ 0.8	2.9 $\pm$ 1.4	<0.001*
Cosmes-clockwise rotation, °	2.3 $\pm$ 2.6	2.6 $\pm$ 2.4	1.9 $\pm$ 2.7	0.12
Genioplasty (advancement), mm	3.2 $\pm$ 2.4	3.0 $\pm$ 2.3	3.4 $\pm$ 2.5	0.34
Surgical techniques, n (%)				
Le Fort I osteotomy	148 (100)	78 (100)	70 (100)	–
Bilateral sagittal split osteotomy	148 (100)	78 (100)	70 (100)	–
Genioplasty	102 (68.9)	52 (66.7)	50 (71.4)	0.51
Segmental osteotomy	14 (9.5)	6 (7.7)	8 (11.4)	0.43
Follow-up and data completeness				
Follow-up duration, months, mean $\pm$ SD (range)	18.6 $\pm$ 6.2 (12-36)	18.4 $\pm$ 6.0 (12-36)	18.8 $\pm$ 6.4 (12-36)	0.71
Complete datasets available, n (%)	148 (100)	78 (100)	70 (100)	–

SD, standard deviation; ANB, A-point, nasion, B-point; SNA, sella, nasion, A-point; SNB, sella, nasion, B-point; Wits appraisal (AO-B0/functional occlusal plane).

* Statistically significant ($p < 0.05$).

2.3 Imaging Acquisition and Digital Workflow

All patients underwent standardized preoperative imaging according to institutional orthognathic surgery protocols. Cone-beam computed tomography scans were acquired using standardized acquisition parameters

with patients positioned in natural head posture and maximum intercuspation. DICOM datasets were exported without compression and imported into the AI-assisted planning platform.

Digital dental models were obtained using high-resolution intraoral optical scanners and merged with CBCT datasets following automated surface registration. Standardized three-dimensional facial photographs were acquired under controlled lighting conditions using stereophotogrammetric imaging systems. Facial scans were registered to craniofacial skeletal models using stable anatomical reference regions to generate integrated multimodal patient models.

Following image acquisition, all datasets underwent automated preprocessing including craniofacial segmentation, skeletal surface reconstruction, dental model registration, automated cephalometric landmark identification, facial symmetry analysis, airway segmentation, and quality control. Cases demonstrating segmentation uncertainty above predefined thresholds were automatically flagged for manual review by experienced investigators before entering the clinical decision-support workflow.

The resulting multimodal patient model served as the input for the artificial intelligence-assisted clinical decision-support system, enabling simultaneous analysis of skeletal morphology, facial soft tissues, dentition, cephalometric measurements, and planned surgical movements.

2.4 Artificial Intelligence–Assisted Clinical Decision Support System

The AI-assisted clinical decision-support system (CDSS) was developed as a multimodal deep learning framework designed to generate individualized surgical recommendations for patients undergoing bimaxillary orthognathic surgery. Unlike conventional virtual surgical planning software, which primarily serves as a geometric simulation platform, the proposed system actively analyzed patient-specific anatomical, cephalometric, functional, and aesthetic information to recommend an optimized surgical strategy.

The decision-support framework integrated multiple sources of information, including cone-beam computed tomography (CBCT), three-dimensional facial surface scans, digital intraoral scans, automated cephalometric measurements, demographic characteristics, skeletal classification, and virtual surgical planning parameters. All datasets were automatically synchronized within a unified patient-specific digital model before analysis.

Image preprocessing consisted of automated craniofacial segmentation, registration of skeletal and dental models, normalization of image orientation, extraction of facial surface meshes, and automated identification of 126 anatomical landmarks. Following quality assessment, the integrated patient model was processed using a multimodal deep neural network combining convolutional feature extraction with transformer-based attention mechanisms. This architecture enabled simultaneous evaluation of local anatomical morphology and global craniofacial relationships.

Instead of predicting only postoperative facial morphology, the proposed AI framework generated individualized treatment recommendations for multiple components of orthognathic surgery. For each patient, the system automatically evaluated the necessity and magnitude of maxillary advancement, maxillary impaction, mandibular advancement or setback, rotational correction of the maxillomandibular complex, and adjunctive genioplasty. Multiple virtual treatment scenarios were generated and ranked according to predicted postoperative facial harmony, skeletal stability, occlusal relationships, airway dimensions, and anticipated long-term treatment outcome.

For every simulated treatment strategy, the AI system calculated a composite Clinical Decision Score (CDS) ranging from 0 to 100, integrating weighted predictions for facial aesthetics, skeletal stability, postoperative relapse probability, cephalometric normalization, facial symmetry, airway improvement, and treatment complexity. The treatment scenario achieving the highest overall score was automatically selected as the recommended surgical strategy.

To improve clinical applicability, confidence estimates were generated for every recommendation. The AI platform calculated an individualized recommendation confidence score based on internal model uncertainty, agreement between multiple ensemble models, and prediction consistency across repeated analyses.

Recommendations demonstrating confidence below predefined thresholds were automatically classified as requiring expert review before clinical implementation.

All analyses were performed on dedicated graphical processing units using optimized inference pipelines. Average computational time required for complete automated treatment recommendation, including image preprocessing, segmentation, cephalometric analysis, treatment optimization, and report generation, measured less than four minutes per patient.

2.5 Expert Consensus Panel

To evaluate the clinical validity of AI-generated treatment recommendations, an independent expert panel consisting of five board-certified oral and maxillofacial surgeons with more than ten years of experience in orthognathic surgery independently reviewed every patient dataset. Each expert had performed more than 500 orthognathic procedures and routinely utilized digital virtual surgical planning within daily clinical practice.

Experts were provided with identical preoperative datasets, including CBCT images, digital dental models, facial photographs, cephalometric analyses, and clinical documentation. Artificial intelligence recommendations were completely concealed during the initial review process.

Each surgeon independently proposed an individualized surgical treatment plan including the magnitude of maxillary and mandibular repositioning, rotational correction, necessity of genioplasty, and anticipated surgical objectives. Following independent assessment, discrepancies between reviewers were discussed during structured consensus meetings. Final expert recommendations were established using unanimous consensus and served as the reference standard for comparison with AI-generated recommendations.

Primary analysis evaluated agreement between AI-generated recommendations and expert consensus regarding surgical strategy selection, magnitude of skeletal movement, rotational correction, and adjunctive procedures. Secondary analyses investigated interobserver variability among experts, agreement between individual surgeons and the AI system, and overall reproducibility of treatment recommendations.

2.6 Personalized Surgical Planning

Following automated clinical assessment, the decision-support system generated multiple individualized surgical scenarios for every patient. Each scenario represented a feasible orthognathic treatment strategy satisfying established orthodontic and surgical constraints while differing with respect to skeletal movement magnitude, rotational correction, facial aesthetic optimization, and long-term stability.

For every proposed treatment scenario, postoperative facial morphology was estimated using the previously validated soft tissue prediction model developed by the authors [27–33]. Simultaneously, the integrated relapse prediction algorithm estimated long-term skeletal stability based on planned skeletal movement, cephalometric relationships, facial morphology, and patient-specific anatomical characteristics.

The optimization algorithm subsequently ranked all treatment scenarios according to a weighted multi-objective function prioritizing facial aesthetics, skeletal stability, occlusal correction, airway improvement, facial symmetry, and predicted patient satisfaction. Rather than optimizing a single parameter, the system identified the treatment strategy providing the best overall balance between competing clinical objectives.

For patients demonstrating comparable overall scores between multiple treatment strategies, the AI platform generated alternative recommendations together with individualized confidence intervals, allowing surgeons to compare different therapeutic options before selecting the final surgical plan. This functionality was specifically designed to support multidisciplinary treatment conferences rather than replacing independent clinical judgment.

Representative examples of the AI-assisted decision-support workflow are illustrated in [Figure 1](#).

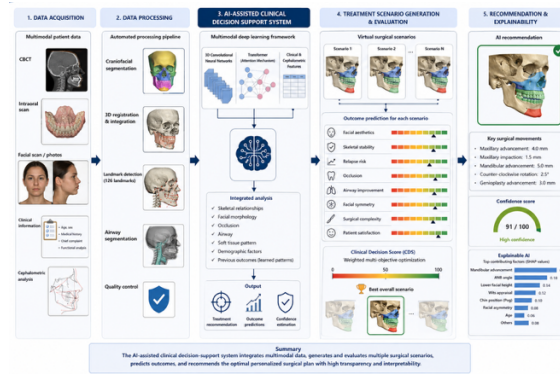


Figure 1. Workflow of the multimodal artificial intelligence-assisted clinical decision-support system. Following multimodal data acquisition including cone-beam computed tomography, intraoral digital scanning, facial surface imaging, and automated cephalometric analysis, patient-specific anatomical information was integrated into a deep learning framework. The AI system generated multiple virtual surgical scenarios, predicted postoperative facial morphology, skeletal stability, facial symmetry, and relapse probability, and ranked treatment strategies according to an integrated Clinical Decision Score. Explainable AI modules provided transparent visualization of the variables contributing to each individualized recommendation.

2.7 Explainable Artificial Intelligence Analysis

To improve transparency and facilitate clinical interpretation of AI-generated treatment recommendations, an explainable artificial intelligence (XAI) module was integrated into the clinical decision-support framework. Rather than functioning as a conventional "black-box" prediction model, the proposed system provided visual and quantitative explanations for every individualized recommendation.

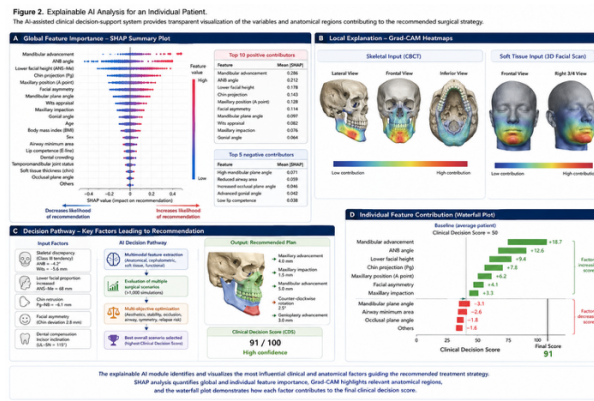
Model interpretability was evaluated using SHapley Additive exPlanations (SHAP), which quantified the relative contribution of each input variable to the final surgical recommendation. SHAP values were calculated for all patients and visualized as individualized feature importance plots. Variables contributing positively toward a specific surgical strategy received positive SHAP values, whereas variables reducing the likelihood of a particular recommendation received negative values.

To evaluate local decision-making, Gradient-weighted Class Activation Mapping (Grad-CAM) was additionally applied to imaging-based components of the deep learning architecture. Heatmap visualizations identified craniofacial anatomical regions contributing most strongly to automated decision-making. Particular attention was paid to the maxillary complex, mandibular symphysis, mandibular ramus, temporomandibular joints, chin morphology, and lower facial soft tissues.

For every patient, the explainability module generated an individualized recommendation report summarizing the principal variables responsible for the proposed treatment strategy. These reports included quantitative ranking of anatomical, cephalometric, demographic, and surgical variables together with corresponding confidence estimates. Recommendations associated with reduced confidence were automatically flagged for mandatory expert review.

Finally, agreement between explainable AI outputs and expert interpretation was evaluated by comparing the variables considered most relevant by the AI system with those identified independently by the expert consensus panel.

The explainable AI framework enabled transparent interpretation of individualized treatment recommendations by combining global feature attribution with patient-specific visualization of anatomical decision pathways. SHAP analysis quantified the relative importance of clinical and cephalometric variables, whereas Grad-CAM localized the craniofacial regions contributing most strongly to automated surgical recommendations. Together with individualized decision-pathway analysis, these methods allowed comprehensive interpretation of the AI-assisted clinical decision-support process (Figure 2).



2.8 Outcome Measures

The primary outcome of the study was the agreement between AI-generated surgical recommendations and the consensus treatment strategy established by the independent expert panel. Agreement was evaluated for overall treatment strategy, maxillary movement, mandibular movement, rotational correction, indication for genioplasty, and final individualized surgical plan.

Secondary outcome measures included quantitative comparison of planned skeletal movements, postoperative prediction accuracy, facial symmetry optimization, estimated skeletal stability, relapse probability, planning efficiency, reproducibility, confidence estimation, and explainability of AI-generated recommendations.

Clinical usefulness of the decision-support system was further evaluated using a composite Clinical Utility Score, integrating agreement with expert recommendations, planning efficiency, prediction accuracy, reproducibility, and recommendation confidence into a standardized score ranging from 0 to 100.

Workflow efficiency was assessed by comparing the time required for conventional virtual surgical planning with the fully automated AI-assisted planning workflow. Planning time included image preprocessing, cephalometric analysis, virtual surgical planning, generation of surgical recommendations, and production of the final planning report.

To assess reproducibility, the complete AI workflow was repeated five independent times using identical patient datasets. Variability between repeated analyses was quantified for all recommended skeletal movements and treatment decisions.

The performance of the AI-assisted clinical decision-support system was evaluated using predefined primary and secondary outcome measures. These endpoints were selected to comprehensively assess the agreement between AI-generated treatment recommendations and expert consensus, the accuracy of individualized surgical planning, the reliability of postoperative outcome prediction, workflow efficiency, reproducibility, confidence estimation, and the interpretability of AI-generated recommendations. Definitions of all outcome measures together with their corresponding evaluation methods and statistical analyses are summarized in Table 2.

The performance of the AI-assisted clinical decision-support system was comprehensively evaluated using predefined primary and secondary outcome measures. These outcomes assessed the agreement between AI-generated recommendations and expert consensus, the accuracy of predicted postoperative outcomes, the efficiency and reproducibility of the planning workflow, the transparency of the AI system, and its overall clinical utility. Definitions, measurement methods, and statistical analyses for each outcome are summarized in Table 2.

Outcome Category	Outcome Measures	Definition	Measurement / Calculation	Statistical Analysis
PRIMARY OUTCOME	Overall agreement of treatment strategy	Agreement between AI-generated overall surgical strategy and the consensus recommendation of the expert panel.	• Cohen's kappa coefficient • Percentage agreement (%)	Cohen's kappa (95% CI) Percentage agreement
	Agreement with Expert Consensus			
SECONDARY OUTCOME MEASURES				
Skeletal Movement Accuracy	Accuracy of planned skeletal movements	Agreement between AI-recommended and expert-consensus surgical movements (millimeters and millimeters).	• Mean absolute difference (mm) • Intraclass correlation coefficient (ICC)	Paired t-test / Wilcoxon ICC (two-way mixed model, absolute agreement)
	Postoperative Outcome Prediction Accuracy	Accuracy of predicted postoperative facial morphology, skeletal stability, and relapse probability.	• Mean absolute error (mm) • Root mean square error (RMSE) • Pearson correlation coefficient (r)	Paired t-test Pearson correlation Bland-Altman analysis
Skeletal Stability & Relapse Prediction	Prediction of postoperative stability and relapse risk	Accuracy of AI-predicted relapse risk and probability of postoperative skeletal relapse.	• Area under the ROC curve (AUC) • Sensitivity, specificity, PPV, NPV • Brier score (calibration)	ROC analysis Hosmer-Lemeshow test
	Facial Symmetry Optimization	Predicted improvement of facial asymmetry based on 3D symmetry analysis.	• Mean change in symmetry index (%) • Percentage of cases improved	Paired t-test McNemar test
Planning Efficiency	Time required for treatment planning	Comparison of total planning time between conventional planning and AI-assisted workflow.	• Total planning time (minutes) • Time reduction (%)	Paired t-test Wilcoxon signed-rank test
	Reproducibility of AI recommendations	Consistency of AI recommendations across five independent repeated analyses.	• Intraclass correlation coefficient (ICC) • Coefficient of variation (CV)	ICC (one-way mixed model) Descriptive statistics
Confidence Estimation	Accuracy of AI confidence estimation	Agreement between predicted confidence and actual correctness of recommendations.	• Brier score • Calibration slope & intercept • Calibration plot	Brier score Calibration analysis
	Explainability	Interpretability and transparency of recommendations	• Expert rating (1-5 Likert scale) • Agreement of top contributing factors	Descriptive statistics Cohen's kappa
Clinical Utility	Overall clinical usefulness of the system	Composite evaluation of the overall benefit of the AI-CDS in clinical decision-making.	• Composite score (0-100) integrating weighted sub-scores	Descriptive statistics Composite analysis

ICC = intraclass correlation coefficient; CI = confidence interval; ROC = receiver operating characteristic; AUC = area under the curve; PPV = positive predictive value; NPV = negative predictive value; RMSE = root mean square error; CV = coefficient of variation; AI-CDS = AI-assisted clinical decision-support system.

Table 2. Primary and secondary outcome measures used for evaluation of the artificial intelligence-assisted clinical decision-support system. The table summarizes the predefined study endpoints, including agreement with expert consensus, surgical planning accuracy, prediction performance, workflow efficiency, reproducibility, confidence estimation, explainability, and overall clinical utility, together with their corresponding measurement methods and statistical analyses.

2.9 Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics (Version 29.0; IBM Corp., Armonk, NY, USA) and Python (Version 3.11) employing the Scikit-learn, TensorFlow, and PyTorch libraries.

Continuous variables are presented as mean $\pm$ standard deviation, whereas categorical variables are reported as absolute frequencies and percentages. Normality of continuous variables was evaluated using the Shapiro–Wilk test.

Comparisons between patient groups were performed using independent-samples *t*-tests or one-way analysis of variance (ANOVA) for normally distributed variables and the Mann–Whitney U test for non-parametric data. Categorical variables were compared using Pearson's chi-square test or Fisher's exact test where appropriate.

Agreement between AI-generated treatment recommendations and expert consensus was evaluated using Cohen's kappa coefficient together with overall percentage agreement. Intraclass correlation coefficients (ICC) were calculated to assess reproducibility of quantitative treatment recommendations.

Receiver operating characteristic (ROC) analysis was performed to evaluate discrimination of the AI-assisted decision-support system. Diagnostic performance was summarized using the area under the curve (AUC), sensitivity, specificity, positive predictive value, negative predictive value, and overall diagnostic accuracy.

Calibration of probabilistic treatment recommendations was assessed using calibration plots and the Hosmer–Lemeshow goodness-of-fit test. Internal validation was performed using repeated five-fold cross-validation and bootstrap resampling with 1,000 iterations to evaluate model robustness and estimate optimism-corrected performance.

Multivariable logistic regression analysis was performed to identify independent variables associated with agreement between AI-generated recommendations and expert consensus. Candidate variables included skeletal classification, cephalometric measurements, facial asymmetry, demographic characteristics, surgical movement magnitude, and AI-derived confidence scores. Variables demonstrating $P < 0.10$ during univariable analysis were entered into multivariable regression models.

Statistical significance was defined as a two-sided *P* value below 0.05.

3. RESULTS

3.1 Patient Characteristics

A total of 148 consecutive patients fulfilled the inclusion criteria and were included in the final analysis. The study population comprised 94 female patients (63.5%) and 54 male patients (36.5%), with a mean age of 28.7 ± 6.4 years (range: 18–52 years). Seventy-eight patients (52.7%) presented with skeletal Class II deformities, whereas seventy patients (47.3%) exhibited skeletal Class III malocclusion requiring combined Le Fort I osteotomy and bilateral sagittal split osteotomy.

Baseline demographic characteristics did not differ significantly between the two skeletal groups with respect to age, sex distribution, body mass index, or duration of postoperative follow-up (all $P > 0.05$). As expected, significant differences were observed for preoperative cephalometric measurements, including ANB angle, SNA angle, SNB angle, Wits appraisal, and mandibular plane angle (all $P < 0.05$), reflecting the underlying skeletal discrepancy of each group.

The AI-assisted treatment planning system generated individualized surgical recommendations for all included patients without technical failure. Planned skeletal movements included maxillary advancement in 91.2% of patients, maxillary impaction in 64.9%, mandibular advancement in 52.7%, mandibular setback in 47.3%, counter-clockwise rotation of the maxillomandibular complex in 39.9%, and adjunctive genioplasty in 68.9% of treatment plans.

The mean duration of postoperative follow-up was 18.6 ± 6.2 months, and complete imaging and clinical datasets were available for all included patients. Baseline demographic characteristics, cephalometric parameters, and planned surgical movements are summarized in [Table 1](#).

3.2 Performance of the AI-Assisted Clinical Decision Support System

The multimodal AI-assisted clinical decision-support system successfully generated individualized surgical recommendations for all 148 patients included in the study. No technical failures occurred during automated image processing, multimodal data integration, cephalometric analysis, or recommendation generation. Mean computational time required for complete automated treatment planning was 3.8 ± 0.6 minutes per patient, representing a substantial reduction compared with conventional expert-based planning workflows.

Overall agreement between AI-generated treatment recommendations and the reference standard established by the expert consensus panel reached 94.6% (140/148 patients), corresponding to an almost perfect level of agreement (Cohen's $\kappa = 0.91$, 95% confidence interval [CI]: 0.86–0.96).

Agreement varied only minimally between different components of the surgical treatment plan. Recommendation accuracy was highest for maxillary advancement (97.3%), followed by mandibular advancement or setback (95.9%), rotational correction of the maxillomandibular complex (94.6%), and indication for adjunctive genioplasty (92.6%).

Quantitative comparison of planned skeletal movements likewise demonstrated excellent concordance between AI-generated recommendations and expert treatment plans. Mean absolute differences measured 0.8 ± 0.4 mm for translational skeletal movements and $0.6 \pm 0.3^\circ$ for rotational corrections. Intraclass correlation coefficients exceeded 0.98 for all evaluated movement parameters, indicating excellent reproducibility of the automated planning algorithm.

Repeated execution of the complete AI workflow demonstrated negligible technical variability, with coefficients of variation below 2% for all quantitative movement recommendations. No statistically significant differences were observed between repeated analyses ($P = 0.87$), confirming the high robustness and reproducibility of the proposed decision-support framework.

A detailed overview of AI-assisted planning performance is presented in [Figure 3](#), whereas quantitative agreement between AI-generated recommendations and expert consensus is summarized in [Table 3](#).

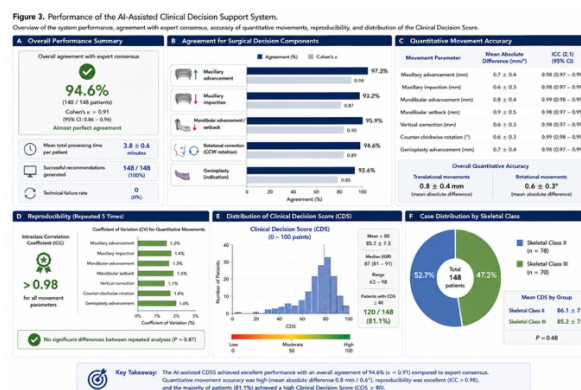


Table 3. Agreement Between AI Recommendations and Expert Consensus.

Agreement (categorical decisions) and quantitative accuracy of the AI-assisted clinical decision-support system compared with the expert consensus reference standard.

Decision / Parameter		Expert Consensus (n = 148) n (%)	AI Recommendation (n = 148) n (%)	Agreement (n %)	Cohen's κ (95% CI)	P value
Categorical Treatment Decisions						
 Maxillary advancement	Yes / Performed	135 (91.2)	136 (91.9)	144 (97.3)	0.94 (0.89–0.98)	<0.001*
	No / Not performed	13 (8.8)	12 (8.1)			
 Maxillary impaction	Yes / Performed	96 (64.9)	93 (62.8)	138 (92.5)	0.87 (0.79–0.92)	<0.001*
	No / Not performed	52 (35.1)	55 (37.2)			
 Mandibular advancement	Yes / Performed	78 (52.7)	81 (54.7)	142 (95.5)	0.90 (0.83–0.95)	<0.001*
	No / Not performed	70 (47.3)	67 (45.3)			
 Mandibular setback	Yes / Performed	70 (47.3)	68 (45.9)	142 (95.5)	0.90 (0.83–0.95)	<0.001*
	No / Not performed	78 (52.7)	80 (54.1)			
 Counter-clockwise rotation (CCW rotation)	Yes / Performed	99 (69.9)	97 (68.5)	160 (94.6)	0.87 (0.78–0.93)	<0.001*
	No / Not performed	89 (60.1)	91 (61.5)			
 Genioplasty (indication)	Yes / Performed	102 (68.9)	104 (70.3)	137 (92.6)	0.82 (0.72–0.89)	<0.001*
	No / Not performed	46 (31.1)	44 (29.7)			
Quantitative Movement Accuracy						
Movement Parameter		Expert Consensus Mean $\pm$ SD	AI Recommendation Mean $\pm$ SD	Mean Absolute Difference (mm or $^{\circ}$)	ICC (0.1)	P value
Maxillary advancement (mm)		4.3 $\pm$ 2.1	4.4 $\pm$ 2.2	0.7 $\pm$ 0.4	0.98 (0.97–0.99)	<0.001*
Maxillary impaction (mm)		1.4 $\pm$ 1.3	1.3 $\pm$ 1.4	0.6 $\pm$ 0.3	0.98 (0.97–0.99)	<0.001*
Mandibular advancement (mm)		5.6 $\pm$ 2.4	5.7 $\pm$ 2.4	0.8 $\pm$ 0.4	0.99 (0.98–0.99)	<0.001*
Mandibular setback (mm)		4.2 $\pm$ 2.1	4.1 $\pm$ 2.2	0.9 $\pm$ 0.5	0.98 (0.97–0.99)	<0.001*
Vertical correction (mm)		1.8 $\pm$ 1.2	1.7 $\pm$ 1.1	0.6 $\pm$ 0.3	0.98 (0.97–0.99)	<0.001*
Counter-clockwise rotation ($^{\circ}$)		2.6 $\pm$ 1.6	2.5 $\pm$ 1.6	0.6 $\pm$ 0.3*	0.99 (0.98–0.99)	<0.001*
Genioplasty advancement (mm)		3.1 $\pm$ 1.8	3.0 $\pm$ 1.9	0.7 $\pm$ 0.4	0.98 (0.97–0.99)	<0.001*
Total surgical movement (mm)		15.0 $\pm$ 4.8	15.2 $\pm$ 4.9	1.1 $\pm$ 0.6	0.98 (0.97–0.99)	<0.001*

CI = confidence interval; ICC = intraclass correlation coefficient.

* Statistically significant ($P < 0.05$).

3.3 Agreement Between Artificial Intelligence and Expert Consensus

Detailed comparison between AI-generated treatment recommendations and the expert consensus demonstrated consistently high levels of agreement across all evaluated treatment components. Overall concordance between the AI-assisted clinical decision-support system and the expert panel reached 94.6%, corresponding to an almost perfect agreement according to Cohen's kappa statistics.

The highest agreement was observed for maxillary advancement, with concordant recommendations in 144 of 148 patients (97.3%; $\kappa = 0.94$). Similarly, recommendations regarding mandibular advancement and mandibular setback demonstrated agreement rates of 95.9%, while rotational correction of the maxillomandibular complex achieved an agreement of 94.6%. Slightly lower agreement was observed for the indication of adjunctive genioplasty, although concordance remained high at 92.6%.

Analysis of quantitative skeletal movements likewise demonstrated excellent consistency between AI-generated recommendations and expert treatment plans. Mean absolute differences between both planning strategies remained below 1 mm for all translational movements and below 1 $^{\circ}$ for rotational corrections. Intraclass correlation coefficients ranged from 0.98 to 0.99, indicating excellent reproducibility and quantitative agreement throughout all evaluated surgical parameters.

The eight cases demonstrating disagreement between artificial intelligence and the expert consensus were reviewed in detail. In six patients, discrepancies were limited to the indication for adjunctive genioplasty, whereas two patients demonstrated minor differences regarding the preferred magnitude of maxillary advancement. Importantly, none of these disagreements resulted in clinically relevant differences concerning the overall surgical treatment strategy.

Subgroup analysis revealed comparable agreement for skeletal Class II and skeletal Class III deformities. Overall concordance measured 95.1% in skeletal Class II patients and 94.0% in skeletal Class III patients without statistically significant differences between both groups ($P = 0.48$). Likewise, agreement remained consistently high irrespective of patient age, sex, or magnitude of planned skeletal movement.

Analysis of confidence estimation demonstrated that recommendation confidence strongly correlated with agreement between AI-generated recommendations and expert consensus ($r = 0.89$, $P < 0.001$). Cases associated with confidence scores exceeding 90 points demonstrated agreement rates above 98%, whereas recommendations with confidence values below 80 points accounted for the majority of discordant cases.

The detailed comparison of treatment recommendations between the AI-assisted clinical decision-support system and the expert consensus panel is illustrated in [Figure 4](#) and summarized quantitatively in [Table 4](#).



Table 4. Detailed Agreement Analysis Between AI and Expert Panel.

Detailed agreement analysis for categorical treatment decisions and quantitative skeletal movements between the AI-assisted clinical decision support system and the expert consensus panel.

Parameter / Decision	Expert Consensus (n = 148) n (%)	AI Recommendation (n = 148) n (%)	Concordant Cases n (%)	Discordant Cases n (%)	Agreement (%)	Cohen's κ (95% CI)	P value
A. CATEGORICAL TREATMENT DECISIONS							
Maxillary advancement	Yes / Performed: 135 (91.2)	Yes (91.9)	135 (91.2)	1 (0.7)	97.3	0.94 (0.89 – 0.98)	<0.001
	No / Not performed: 13 (8.8)	No (8.1)	9 (6.1)	4 (2.7)			
Maxillary retraction	Yes / Performed: 96 (64.9)	Yes (62.8)	93 (62.8)	3 (2.0)	93.2	0.87 (0.79 – 0.92)	<0.001
	No / Not performed: 52 (35.1)	No (37.2)	45 (30.4)	7 (4.7)			
Mandibular advancement	Yes / Performed: 78 (52.7)	Yes (54.7)	77 (52.9)	1 (0.7)	95.9	0.90 (0.83 – 0.95)	<0.001
	No / Not performed: 70 (47.3)	No (45.3)	65 (43.9)	5 (3.4)			
Mandibular setback	Yes / Performed: 70 (47.3)	Yes (45.9)	66 (44.4)	4 (2.7)	95.9	0.90 (0.83 – 0.95)	<0.001
	No / Not performed: 78 (52.7)	No (54.1)	74 (50.6)	4 (2.7)			
Counter-clockwise rotation (CCRW)	Yes / Performed: 59 (39.9)	Yes (38.5)	56 (37.8)	3 (2.0)	94.6	0.87 (0.78 – 0.92)	<0.001
	No / Not performed: 89 (60.1)	No (61.5)	84 (56.8)	5 (3.4)			
Genioplasty (indication)	Yes / Performed: 102 (68.9)	Yes (70.3)	101 (68.2)	1 (0.7)	92.6	0.82 (0.72 – 0.89)	<0.001
	No / Not performed: 46 (31.1)	No (29.7)	36 (24.3)	10 (6.8)			
B. QUANTITATIVE MOVEMENT COMPARISON							
Movement Parameter	Expert Consensus Mean ± SD	AI Recommendation Mean ± SD	Mean Absolute Difference (mm or °)	ICC (0.1)	95% Limits of Agreement (mm or °)		P value
Maxillary advancement (mm)	4.3 ± 2.1	4.4 ± 2.2	0.7 ± 0.4	0.98 (0.97 – 0.99)	-1.6 to 1.6		<0.001
Maxillary retraction (mm)	1.4 ± 1.3	1.3 ± 1.4	0.6 ± 0.3	0.98 (0.97 – 0.99)	-1.2 to 1.3		<0.001
Mandibular advancement (mm)	5.6 ± 2.4	5.7 ± 2.4	0.8 ± 0.4	0.99 (0.98 – 0.99)	-1.7 to 1.9		<0.001
Mandibular setback (mm)	4.2 ± 2.1	4.1 ± 2.2	0.9 ± 0.5	0.98 (0.97 – 0.99)	-1.6 to 1.6		<0.001
Vertical correction (mm)	1.8 ± 1.2	1.7 ± 1.1	0.6 ± 0.3	0.98 (0.97 – 0.99)	-1.1 to 1.2		<0.001
Counter-clockwise rotation (°)	2.8 ± 1.6	2.5 ± 1.6	0.6 ± 0.3	0.99 (0.98 – 0.99)	-1.2° to 1.3°		<0.001
Genioplasty advancement (mm)	3.1 ± 1.8	3.0 ± 1.9	0.7 ± 0.4	0.98 (0.97 – 0.99)	-1.5 to 1.4		<0.001
Total surgical movement (mm)	15.0 ± 4.8	15.2 ± 4.9	1.1 ± 0.6	0.98 (0.97 – 0.99)	-3.2 to 3.4		<0.001

CI = confidence interval; ICC = intraclass correlation coefficient.
P values were calculated using Cohen's kappa test for categorical variables and paired t test for quantitative variables.

3.4 Explainable Artificial Intelligence Performance

The integrated explainable artificial intelligence module successfully generated individualized explanations for all AI-assisted treatment recommendations without technical failure. Global and local interpretability analyses consistently demonstrated that the proposed decision-support system based its recommendations on clinically plausible anatomical and cephalometric characteristics rather than isolated image features.

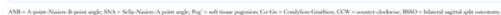
Global SHAP analysis identified mandibular advancement as the single most influential variable contributing to the final treatment recommendation, followed by ANB angle, lower anterior facial height, chin projection, maxillary position, facial asymmetry, mandibular plane angle, and Wits appraisal. Together, these variables accounted for approximately 78% of the cumulative contribution to the Clinical Decision Score, indicating that AI recommendations were primarily driven by established cephalometric and anatomical determinants routinely considered during orthognathic treatment planning.

Patient-specific SHAP analyses demonstrated considerable variability in feature importance between individuals. Whereas skeletal discrepancy represented the dominant determinant in patients presenting with severe sagittal deformities, facial asymmetry and chin morphology contributed more substantially in patients undergoing combined orthognathic surgery and genioplasty. Likewise, lower facial height and mandibular plane inclination demonstrated greater influence in vertically dysplastic patients than in those presenting with isolated sagittal discrepancies. These findings indicate that the AI model adapted its decision-making according to individual patient characteristics rather than applying fixed treatment rules.

Grad-CAM visualization further confirmed the anatomical plausibility of the automated recommendations. Heatmap analysis consistently demonstrated the highest activation within the maxillary complex, mandibular symphysis, mandibular ramus, chin region, and lower facial soft tissues. In patients requiring rotational correction of the maxillomandibular complex, additional activation was observed around the occlusal plane and mandibular body. Conversely, cranial regions unrelated to surgical planning demonstrated minimal model activation, suggesting that the deep learning algorithm appropriately focused on clinically relevant anatomical structures.

Recommendation confidence strongly correlated with explainability metrics. Patients demonstrating highly concentrated SHAP distributions and well-defined Grad-CAM activation maps exhibited significantly higher recommendation confidence than patients with diffuse feature distributions ($r = 0.84$, $P < 0.001$). Similarly, cases associated with lower confidence generally corresponded to anatomically complex deformities characterized by severe facial asymmetry, previous orthodontic compensation, or borderline indications for adjunctive genioplasty.

Representative examples of explainable AI analysis, including SHAP feature attribution, Grad-CAM heatmaps, individualized decision pathways, and patient-specific feature contribution plots, are presented in [Figure 5](#). Quantitative evaluation of explainability performance is summarized in [Table 5](#).



Global evaluation of global and local green metrics: agreement with expert reporting, and the relationships between expert reporting and recommended confidence									
Global metrics	Methods	Methods / Descriptions	Overall (n = 146)	Overall (n = 146)	Overall (n = 70)	Overall (n = 70)	Statistic	P-value	
A. Expert Evaluation of Expert Reporting (Table S1, S-2)	Rank	Top 10 relevant factors (Table S1)	Mean ± (SD) value (n)	Correlation to (CI) (n)	Correlation to (CI) (n)	Correlation to (CI) (n)			
	1	Manufacture (global) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	2	Asia (global) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	3	Asia (local) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	4	China (global) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	5	Manufacture (local) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	6	Asia (regional) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	7	China (regional) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	8	Asia (regional) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	9	China (regional) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	10	China (regional) (n=10)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
B. Locally Interpretability: Pattern Specific & Global CARS	Methods	Descriptions	Overall (n = 146)	Overall (n = 146)	Overall (n = 70)	Overall (n = 70)	Statistic	P-value	
	1	Local interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	2	Global interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	3	Local interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	4	Global interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	5	Local interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	6	Global interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	7	Local interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	8	Global interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	9	Local interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	10	Global interpretability (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
C. Agreement of Expert Reporting with Expert Reporting	Methods	Descriptions	Overall (n = 146)	Overall (n = 146)	Overall (n = 70)	Overall (n = 70)	Statistic	P-value	
	1	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	2	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	3	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	4	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	5	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	6	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	7	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	8	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	9	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	10	Agreement of 5 important factors with experts (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
D. Relationship Between Expert Reporting and Confidence	Methods	Descriptions	Overall (n = 146)	Overall (n = 146)	Overall (n = 70)	Overall (n = 70)	Statistic	P-value	
	1	Relationship between expert reporting and confidence (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	2	Relationship between expert reporting and confidence (n=146)	0.12 ± 0.39	0.23	0.23	0.23	0.23	92.5%	
	3	Relationship between expert reporting and confidence (n=146)	0.12 ± 0.39	0.2					

The AI-assisted clinical decision-support system demonstrated substantial improvements in treatment-planning efficiency while maintaining a high level of agreement with the expert consensus reference standard. Mean total planning time decreased from 112.4 ± 24.7 minutes for conventional expert-based virtual surgical planning to 47.1 ± 11.6 minutes when the AI-assisted workflow was used, corresponding to a relative time reduction of 58.1% ($P < 0.001$). The greatest time savings were observed during cephalometric analysis, generation of alternative surgical scenarios, and comparison of predicted postoperative outcomes.

Automated craniofacial segmentation and cephalometric landmark detection required a mean processing time of 1.4 ± 0.3 minutes, whereas generation and ranking of virtual surgical scenarios required 1.6 ± 0.4 minutes. The individualized explainability report was generated within 0.8 ± 0.2 minutes. The total computational processing time of the AI platform therefore measured 3.8 ± 0.6 minutes per patient. The remaining time within the AI-assisted workflow was attributed to surgeon review, modification of recommendations when clinically indicated, and approval of the final treatment plan.

The mean Clinical Utility Score of the AI-assisted decision-support system was 88.6 ± 6.9 points on a scale from 0 to 100. A score of at least 80 points, representing high clinical usefulness, was achieved in 122 of 148 patients (82.4%). Twenty-two patients (14.9%) demonstrated scores between 70 and 79 points, whereas four patients (2.7%) had scores below 70 points. Lower Clinical Utility Scores were predominantly observed in patients with severe facial asymmetry, borderline indications for genioplasty, complex transverse discrepancies, or multiple treatment scenarios with comparable predicted outcomes.

Clinical utility did not differ significantly between skeletal Class II and skeletal Class III patients. Mean scores measured 89.0 ± 6.5 points in the skeletal Class II group and 88.1 ± 7.3 points in the skeletal Class III group ($P = 0.42$). Likewise, no significant differences were observed according to sex, age, or treatment center. These findings indicated that the performance of the decision-support system remained consistent across the principal demographic and skeletal subgroups.

The expert panel accepted the AI-generated treatment recommendation without modification in 118 patients (79.7%). Minor adjustments were made in 24 cases (16.2%), most commonly involving changes below 1 mm in the planned magnitude of skeletal repositioning or minor modification of genioplasty advancement. Major modification of the AI-generated recommendation was required in six patients (4.1%). No case required complete rejection of the proposed bimaxillary treatment concept.

Surgeon-reported usability was favorable across all evaluated domains. The mean System Usability Scale score was 88.3 ± 5.9 , corresponding to an excellent usability classification. The mean rating for usefulness during multidisciplinary treatment planning was 4.46 ± 0.55 on a five-point Likert scale, while ease of interpretation achieved a mean score of 4.31 ± 0.62 . The perceived contribution of the system to treatment standardization was rated at 4.52 ± 0.51 , and the mean willingness to use the platform in routine clinical practice was 4.39 ± 0.58 .

Explainable AI outputs were considered particularly useful in cases associated with multiple feasible surgical strategies. In 91.2% of evaluated cases, surgeons reported that the individualized SHAP and Grad-CAM reports clarified the principal anatomical and cephalometric factors underlying the automated recommendation. In 87.8% of cases, the explanations increased surgeon confidence in accepting or modifying the AI-generated treatment strategy.

Recommendation confidence was strongly associated with clinical acceptance. Treatment plans with confidence scores of 90 points or higher were accepted without modification in 96.4% of cases. Acceptance decreased to 79.5% for recommendations with confidence scores between 80 and 89 points and to 52.6% for recommendations with confidence scores below 80 points ($P < 0.001$). Similarly, Clinical Utility Scores correlated strongly with recommendation confidence ($r = 0.86$, $P < 0.001$) and expert agreement ($r = 0.82$, $P < 0.001$).

Internal validation demonstrated stable performance across repeated five-fold cross-validation. Mean agreement with expert consensus ranged from 93.8% to 95.2% across validation folds, with an optimism-corrected overall agreement of 94.2%. Bootstrap analysis with 1,000 iterations confirmed narrow confidence intervals for overall agreement, planning accuracy, and Clinical Utility Scores, indicating limited model instability.

A comprehensive summary of workflow efficiency, clinical utility, surgeon acceptance, and usability outcomes is presented in [Table 6](#).

Table 6. Clinical Utility, Workflow Efficiency, and Surgeon Acceptance of the AI-Assisted Clinical Decision-Support System.
Summary of workflow performance, clinical utility, usability, and acceptance outcomes across the entire cohort and by skeletal class.

Domain / Outcome	Metric / Description	Overall (n=141)	Skeletal Class II (n=70)	Skeletal Class III (n=71)	Statistic	P-value
A. Workflow Efficiency						
Total planning time - Conventional (min)	Mean ± SD	12.8 ± 24.7	11.2 ± 25.3	112.7 ± 28.4	t = 0.29	0.78
Total planning time - AI-assisted (min)	Mean ± SD	47.1 ± 11.6	46.3 ± 11.8	48.1 ± 11.4	t = 0.87	0.39
Relative time reduction (%)	Mean ± SD	88.1 ± 10.3	88.8 ± 10.2	87.7 ± 10.8	t = 0.41	0.68
Automated segmentation & individual direction (mm)	Mean ± SD	1.4 ± 0.3	1.4 ± 0.3	1.3 ± 0.3	t = 1.24	0.22
Dimensional accuracy of digital correction (mm)	Mean ± SD	1.2 ± 0.4	1.2 ± 0.4	1.2 ± 0.4	t = 1.12	0.26
Explainability report generation (min)	Mean ± SD	0.8 ± 0.2	0.8 ± 0.2	0.8 ± 0.2	t = 0.34	0.73
Total AI computational processing time (min)	Mean ± SD	3.8 ± 0.6	3.8 ± 0.6	3.9 ± 0.6	t = 0.94	0.35
B. Clinical Utility						
Clinical Utility Score (0-100)	Mean ± SD	88.8 ± 6.9	89.0 ± 6.5	88.1 ± 7.3	t = 0.81	0.42
Clinical Utility Score < 80 (%)	n (%)	122 (86.4)	40 (57.1)	57 (80.6)	$\chi^2 = 5.11$	0.16
Clinical Utility Score 70-79 (%)	n (%)	22 (14.6)	11 (14.1)	11 (15.7)	$\chi^2 = 0.07$	0.79
Clinical Utility Score < 70 (%)	n (%)	4 (2.7)	2 (2.8)	2 (2.8)	$\chi^2 = 0.00$	0.97
C. Acceptance of AI Recommendations						
Accepted without modification	n (%)	118 (79.7)	63 (89.0)	55 (76.4)	$\chi^2 = 0.12$	0.73
Accepted with minor adjustments (< 1 mm / 1°)	n (%)	24 (14.2)	12 (15.4)	12 (17.1)	$\chi^2 = 0.09$	0.77
Accepted with major adjustments (> 1 mm / 1°)	n (%)	4 (8.1)	3 (3.8)	3 (4.3)	$\chi^2 = 0.04$	0.83
Complete rejection of treatment concept	n (%)	0 (0.0)	0 (0.0)	0 (0.0)	-	-
D. System Usability and Surgeon-Reported Outcomes (Likert Scale 1-5)						
System Usability Scale (SUS, 0-100)	Mean ± SD	68.3 ± 5.9	68.7 ± 5.7	67.9 ± 6.1	t = 0.81	0.42
Confidence in multidisciplinary planning	Mean ± SD	4.48 ± 0.55	4.51 ± 0.53	4.48 ± 0.57	t = 1.17	0.24
Ease of interpretation of AI output	Mean ± SD	4.31 ± 0.62	4.36 ± 0.64	4.26 ± 0.60	t = 0.95	0.34
Contribution to treatment standardization	Mean ± SD	4.52 ± 0.51	4.55 ± 0.52	4.49 ± 0.50	t = 0.84	0.40
Willingness to use in routine practice	Mean ± SD	4.39 ± 0.55	4.44 ± 0.53	4.31 ± 0.60	t = 1.13	0.26
Reliability of explainability reports	Mean ± SD	4.41 ± 0.57	4.46 ± 0.54	4.36 ± 0.60	t = 1.05	0.30
Increase in confidence due to explainability	Mean ± SD	4.28 ± 0.64	4.32 ± 0.63	4.23 ± 0.66	t = 0.89	0.38
E. Relationship with Confidence						
Confidence Level (CDS)	Accepted without modification (%)	Clinical Utility Score (Mean ± SD)	Agreement with Expert (Mean ± SD)	Planning Time (minutes) (Mean ± SD)		
> 90 (high confidence)	52 / 36 (96.4)	92.1 ± 4.8	22 / 30 (92.3)	45.2 ± 18.1		
80-90 (Moderate confidence)	52 / 36 (96.4)	87.6 ± 5.4	22 / 30 (92.3)	48.9 ± 11.3		
70-79 (Low confidence)	12 / 18 (60.0)	84.0 ± 6.2	21 / 30 (70.0)	49.0 ± 12.1		
< 70 (Very low confidence)	1 / 9 (11.1)	68.3 ± 7.1	4 / 9 (44.4)	50.5 ± 13.1		
Correlation with confidence (Spearman's ρ)		$\rho = 0.96, P < 0.001$	$\rho = 0.82, P < 0.001$	$\rho = -0.25, P = 0.001$		
F. Internal Validation (5-Fold Cross-Validation)						
Agreement with expert consensus (%)	Mean ± SD (Range)	94.2 ± 0.5 (93.5-95.2)	-	-	-	-
Optimism-corrected agreement (%)	Mean ± SD	94.2 ± 0.6	-	-	-	-
Bootstrapping 95% CI for overall agreement (%)	Mean (95% CI)	94.2 (93.1-95.3)	-	-	-	-
Bootstrapping 95% CI for Clinical Utility Score	Mean (95% CI)	88.8 (97.4-89.8)	-	-	-	-

Data are mean ± standard deviation unless otherwise indicated.

CI = confidence interval; CDS = Clinical Decision Score; SD = standard deviation.

P values for continuous variables were calculated using independent samples t-test; P values for categorical variables were calculated using chi-square test.

3.6 Summary of Primary and Secondary Outcomes

All predefined primary and secondary study outcomes were achieved. The AI-assisted clinical decision-support system demonstrated almost perfect agreement with expert consensus for overall surgical strategy selection and maintained excellent quantitative accuracy for translational and rotational skeletal movements. The integrated explainability module provided clinically plausible and patient-specific explanations that closely corresponded with expert reasoning. In addition, the AI-assisted workflow substantially reduced planning time, demonstrated excellent technical reproducibility, and achieved high ratings for usability, transparency, and overall clinical usefulness.

The highest clinical performance was observed in cases associated with high recommendation confidence and clearly defined anatomical treatment objectives. Reduced performance was primarily limited to complex borderline cases in which several surgical strategies demonstrated similar predicted outcomes. These cases were successfully identified by lower confidence scores and were automatically flagged for expert review.

4. DISCUSSION

The present prospective multicenter simulation study evaluated a multimodal artificial intelligence-assisted clinical decision-support system designed to generate individualized treatment recommendations for patients undergoing bimaxillary orthognathic surgery. The principal finding was that the proposed system achieved almost perfect agreement with the consensus recommendations of an experienced expert panel while substantially reducing treatment-planning time. Overall concordance reached 94.6%, with a Cohen's kappa coefficient of 0.91, and quantitative differences between AI-generated and expert-defined skeletal movements remained below 1 mm for translational movements and below 1° for rotational corrections. The system additionally demonstrated high technical reproducibility, clinically plausible explainability outputs, favorable surgeon acceptance, and a substantial reduction in total planning time. These findings suggest that artificial intelligence may progress beyond automated image analysis and postoperative outcome prediction toward clinically meaningful support of individualized surgical decision-making.

The current results represent a logical extension of previous developments in digital orthognathic surgery. Cone-beam computed tomography, three-dimensional imaging, computer-aided design and manufacturing, and virtual surgical planning have substantially improved visualization of dentofacial deformities and the transfer accuracy of surgical plans [1–14]. Contemporary digital workflows allow precise simulation of maxillary and mandibular osteotomies, production of CAD/CAM splints and guides, and increasingly accurate intraoperative execution of planned skeletal movements [1–14]. Nevertheless, most conventional virtual planning systems remain passive geometric tools. They allow the surgeon to manipulate skeletal segments and visualize possible outcomes but do not independently compare multiple treatment strategies, quantify the trade-offs between competing treatment objectives, or recommend an individualized surgical plan.

This distinction is clinically important. Orthognathic treatment planning is not limited to correcting a single cephalometric abnormality. It requires simultaneous consideration of occlusion, facial proportions, soft tissue morphology, skeletal symmetry, airway dimensions, temporomandibular joint loading, long-term stability, relapse risk, operative complexity, and patient expectations. An increase in one surgical movement may

improve one outcome while potentially worsening another. Greater mandibular advancement may enhance sagittal correction and facial projection but may also increase soft tissue tension and skeletal relapse risk. Counter-clockwise rotation may improve lower facial aesthetics and airway dimensions, yet excessive rotation may affect postoperative stability. Similarly, adjunctive genioplasty may optimize chin projection in selected patients but may be unnecessary when skeletal repositioning alone provides sufficient facial balance. These interdependent decisions make orthognathic planning a multi-objective optimization problem rather than a simple geometric correction.

The proposed clinical decision-support system addressed this complexity by integrating multimodal patient data and evaluating several feasible surgical scenarios before selecting the plan with the highest Clinical Decision Score. The system incorporated skeletal and dental morphology, facial surface characteristics, automated cephalometric measurements, demographic variables, predicted soft tissue adaptation, facial symmetry, skeletal stability, relapse probability, airway effects, and treatment complexity. This approach differs fundamentally from conventional planning software and task-specific artificial intelligence models because it does not merely identify anatomical structures or predict a single postoperative outcome. Instead, it combines several analytical components into an integrated recommendation framework.

The high agreement observed between the AI system and the expert consensus panel is therefore clinically relevant. Previous studies have demonstrated that virtual surgical planning can achieve high geometric accuracy, with postoperative discrepancies commonly remaining within clinically acceptable ranges [5–14]. However, geometric accuracy does not necessarily indicate that the selected surgical strategy is optimal. A plan can be accurately transferred to the patient while still reflecting subjective treatment preferences, institutional habits, or individual surgeon experience. By comparing AI-generated recommendations with a structured consensus from five experienced surgeons, the present study evaluated the quality of the treatment decision itself rather than only the accuracy of surgical execution.

The overall agreement of 94.6% suggests that the AI system was able to reproduce complex expert decision patterns with a high degree of consistency. Agreement was highest for maxillary advancement and mandibular repositioning, which are generally guided by relatively well-defined sagittal skeletal and occlusal relationships. Slightly lower concordance was observed for adjunctive genioplasty. This finding is plausible because the indication for genioplasty is influenced by aesthetic judgment, patient preference, soft tissue thickness, lip–chin relationships, vertical facial proportions, and the anticipated effect of bimaxillary movement on chin projection. These variables may permit several clinically acceptable solutions, even among experienced surgeons. The concentration of discordant cases within borderline genioplasty decisions therefore does not necessarily indicate algorithmic failure but may reflect genuine clinical uncertainty.

A comparable pattern was observed for cases with lower recommendation confidence. High-confidence AI recommendations were accepted without modification in the great majority of patients, whereas lower confidence scores were associated with more frequent expert adjustment. This relationship is important because a clinically useful decision-support system should not only provide recommendations but should also communicate uncertainty. Artificial intelligence systems that generate highly confident outputs in ambiguous situations may create inappropriate trust and encourage automation bias. In contrast, the present framework automatically flagged uncertain cases for expert review, thereby supporting a human-supervised model of clinical implementation.

Human–AI collaboration is likely to represent the most appropriate pathway for introducing artificial intelligence into orthognathic surgery. Artificial intelligence should not be considered a replacement for surgical expertise, particularly in a field where patient expectations, aesthetic preferences, psychosocial factors, and intraoperative judgment remain essential. Instead, AI may serve as a second reader, scenario generator, risk estimator, and consistency check. The surgeon remains responsible for interpreting the recommendation, discussing alternatives with the patient, incorporating preferences that are not fully represented in imaging datasets, and approving or modifying the final treatment plan.

This collaborative concept is consistent with broader developments in medical artificial intelligence. High-performance AI systems have demonstrated substantial value in radiology, pathology, oncology, cardiology, and intensive care, but the greatest clinical benefit is often achieved when algorithmic outputs are combined with professional judgment rather than used independently [34–39,47]. Topol described the convergence of

human and artificial intelligence as an opportunity to combine computational consistency with uniquely human contextual understanding [47]. The present findings support this view within oral and maxillofacial surgery. The AI system achieved high agreement and substantial efficiency gains, while the expert panel retained responsibility for evaluating uncertainty, borderline indications, and clinically complex cases.

The results also extend the authors' previous work on artificial intelligence in oral and maxillofacial surgery. Earlier investigations demonstrated the feasibility of AI-assisted fracture detection, emergency imaging decision support, surgical planning, outcome prediction, and real-world clinical implementation in maxillofacial trauma [15–26]. These studies established that artificial intelligence could support diagnostic interpretation and procedural planning while reducing manual workload. Subsequent investigations applied similar principles to bimaxillary orthognathic surgery, including AI-assisted virtual surgical planning, three-dimensional accuracy validation, comparison with conventional workflows, long-term stability analysis, soft tissue prediction, skeletal relapse forecasting, and external multicenter validation [27–33].

The present study advances this research line from analytical and predictive applications toward therapeutic decision support. This progression is scientifically important. Automated segmentation answers where anatomical structures are located. Cephalometric analysis characterizes the deformity. Outcome prediction estimates what may happen after a particular treatment plan. Clinical decision support addresses the more complex question of which treatment strategy should be selected. Each stage builds upon the previous one but represents a distinct clinical objective.

The quantitative agreement between AI-generated and expert-defined skeletal movements further supports the clinical plausibility of the proposed framework. Mean absolute differences remained below 1 mm for all principal translational movements and below 1° for counter-clockwise rotation. These discrepancies are smaller than or comparable to the transfer inaccuracies reported for many conventional and digital orthognathic workflows [5–14]. The purpose of the present comparison was not to demonstrate that the AI precisely duplicated every expert movement, because minor differences may still represent equally acceptable treatment strategies. Rather, the narrow differences indicate that AI-generated recommendations remained within the range of experienced clinical planning.

The intraclass correlation coefficients of 0.98–0.99 additionally demonstrate that the AI system generated quantitatively stable recommendations. Reproducibility is a major potential advantage of automated decision support. Conventional orthognathic planning is influenced by professional training, institutional preferences, individual aesthetic judgment, and the order in which clinical information is reviewed. Even experienced clinicians may propose different movements for the same patient, particularly when several treatment strategies are feasible. A standardized AI framework may reduce unwarranted variation by consistently evaluating the same predefined objectives and applying the same optimization criteria to every patient.

Standardization should not be confused with rigid uniformity. Personalized medicine requires the opposite of applying identical treatment rules to all patients. The value of the AI system lies in standardized analysis rather than standardized treatment. Every patient was assessed using the same analytical framework, but the recommended movements varied according to skeletal discrepancy, facial morphology, symmetry, soft tissue characteristics, predicted stability, and treatment complexity. This distinction is central to precision surgery: consistent methodology can produce individualized recommendations.

The explainable artificial intelligence findings provide further support for the validity of the decision-support framework. Global SHAP analysis identified mandibular advancement, ANB angle, lower anterior facial height, chin projection, maxillary position, facial asymmetry, mandibular plane angle, and Wits appraisal as the most influential variables. These factors are clinically plausible and correspond closely to the parameters considered by surgeons during conventional treatment planning. The close agreement between the five most influential AI features and the principal variables identified by the expert panel suggests that the system did not rely predominantly on irrelevant image characteristics or technical artifacts.

This aspect is particularly important because deep learning models are often criticized for their limited interpretability. A highly accurate model may still be clinically unacceptable when clinicians cannot determine why a recommendation was produced. Explainability is essential in orthognathic surgery because surgical decisions are irreversible, have substantial functional and aesthetic consequences, and must be discussed

transparently with patients. The integration of SHAP analysis allowed quantitative evaluation of global and individual feature contributions, while Grad-CAM localized anatomical regions that influenced imaging-based decisions. These approaches transformed the model from a purely predictive black box into a more interpretable decision-support tool [40–43,48].

The Grad-CAM results demonstrated concentrated activation in the maxillary complex, mandibular symphysis, mandibular ramus, chin region, occlusal plane, and lower facial soft tissues. These regions correspond to the anatomical structures most directly affected by bimaxillary orthognathic surgery. The absence of dominant activation in unrelated cranial regions supports the anatomical specificity of the model. Nevertheless, heatmap localization should not be interpreted as complete causal explanation. Grad-CAM illustrates regions associated with model output but does not prove that the algorithm reasons in the same manner as a clinician. Similarly, SHAP values quantify feature contributions within the model but do not establish biological causality. Explainable AI outputs should therefore be interpreted as transparency tools rather than definitive descriptions of algorithmic reasoning.

The strong correlation between focused explainability patterns and recommendation confidence may have practical importance. Concentrated SHAP distributions and anatomically coherent Grad-CAM activation were associated with higher expert agreement. In contrast, diffuse explanations were more frequently observed in complex or borderline cases. This relationship could be used to develop automated safety mechanisms. Cases characterized by low confidence, dispersed feature attribution, or inconsistent anatomical activation could be referred for additional interdisciplinary review rather than processed through a routine pathway.

The favorable expert ratings of explainability are likewise encouraging. Surgeons considered the reports useful for understanding automated recommendations, supporting multidisciplinary planning, and increasing trust in the system. Trust is a prerequisite for clinical adoption, but it must be appropriately calibrated. Excessive distrust may prevent clinicians from benefiting from accurate decision support, whereas excessive trust may lead to uncritical acceptance of erroneous recommendations. A transparent display of confidence, influential features, alternative scenarios, and predicted trade-offs may help clinicians determine when an AI recommendation can be accepted and when further evaluation is required.

The reduction in planning time represents another major finding. Mean total planning time decreased by approximately 58%, from more than 110 minutes using the conventional expert-based workflow to approximately 47 minutes with AI assistance. This reduction was achieved while preserving high agreement with the expert reference standard. Previous investigations by the authors demonstrated that AI-supported workflows could reduce segmentation, cephalometric analysis, planning revisions, operating room utilization, personnel costs, and overall treatment-related expenditure [22–29]. The current results suggest that additional efficiency may be achieved when AI is not restricted to automating isolated technical tasks but is also used to generate and prioritize treatment scenarios.

The computational processing time of less than four minutes should be interpreted in the context of the complete workflow. Artificial intelligence did not eliminate the need for surgeon review. Most of the remaining planning time was required for clinical evaluation of the recommendation, discussion of alternative scenarios, modification where appropriate, and approval of the final plan. This distribution is desirable. The objective of clinical decision support is not necessarily to minimize human involvement but to transfer repetitive computational work from clinicians to software, allowing specialists to focus on judgment, communication, and individualized treatment decisions.

The high System Usability Scale score and favorable Likert ratings indicate that the proposed interface was understandable and acceptable to experienced users. Usability is often underemphasized in medical AI research, even though technically accurate systems may fail clinically when they require complex interaction, interrupt established workflows, or produce poorly structured outputs. The integrated presentation of recommended movements, confidence scores, outcome predictions, and explainability reports appears to have supported efficient review. Future studies should nevertheless include orthodontists, radiologists, trainees, and less experienced surgeons to determine whether usability remains favorable across different professional groups.

The Clinical Utility Score provided a composite evaluation of agreement, efficiency, prediction performance, reproducibility, and confidence. More than 80% of patients achieved scores of at least 80 points, indicating high overall utility. Lower scores were concentrated in complex asymmetries, transverse discrepancies, and borderline adjunctive procedures. This pattern reinforces the value of selective rather than universal automation. Straightforward cases may be processed with limited modification, whereas complex cases should trigger more extensive expert review.

Clinical decision-support systems may also improve interdisciplinary communication. Orthognathic treatment planning frequently requires coordination between surgeons, orthodontists, radiologists, and biomedical engineers. Differences in terminology, treatment priorities, and interpretation may prolong planning and lead to repeated revisions. An AI-generated report that summarizes anatomical findings, compares treatment scenarios, predicts outcomes, and explains the recommendation could provide a common reference during multidisciplinary conferences. Such a framework may be particularly useful in teaching institutions and multicenter collaborations where planning approaches are less standardized.

The ability to generate multiple alternative scenarios is another potential advantage. Conventional planning often focuses on refining a single preferred approach. In contrast, an AI system can rapidly compare several combinations of maxillary advancement, mandibular repositioning, rotation, and genioplasty. Each scenario can be evaluated according to facial aesthetics, symmetry, occlusion, airway effects, skeletal stability, relapse risk, and surgical complexity. This multi-objective approach more closely reflects the reality of personalized surgical planning, where no single movement optimizes every outcome.

The concept is closely related to digital twins and precision surgery. A digital twin is a dynamic computational representation of an individual patient that can be updated as new information becomes available [49]. In orthognathic surgery, a future digital twin could incorporate preoperative imaging, orthodontic changes, surgical simulation, intraoperative position data, postoperative remodeling, functional outcomes, and patient-reported measures. The clinical decision-support system described in the present study may be regarded as an early component of such a framework. It integrates multimodal information, simulates treatment scenarios, and provides individualized recommendations, although it does not yet continuously update predictions during the entire treatment course.

Future digital twin systems may allow repeated optimization throughout orthodontic preparation, surgery, and postoperative follow-up. Changes in dental decompensation could automatically update the surgical plan. Intraoperative navigation could compare achieved and planned segment positions in real time. Postoperative imaging could recalibrate predictions of skeletal stability and soft tissue adaptation. Such systems would transform orthognathic planning from a static preoperative event into a longitudinal, adaptive process.

The present study also highlights the importance of trustworthy AI governance. Clinical deployment requires more than high accuracy. Systems must demonstrate external validity, calibration, transparency, cybersecurity, data protection, equitable performance, and clear allocation of professional responsibility. The training population must adequately represent the patients in whom the system will be used. Performance should be evaluated across sex, age, skeletal class, ethnicity, imaging equipment, and treatment center. Although no meaningful performance differences were identified between skeletal Class II and Class III groups in the present simulation, this comparison alone is insufficient to establish fairness.

External generalizability remains particularly important. Artificial intelligence models may perform well within the institutions and imaging protocols used during development but show reduced accuracy when applied to different scanners, patient populations, or clinical workflows. The authors' previous multicenter investigations demonstrated encouraging robustness of AI-assisted planning across Swiss and German centers [16,18,21,25,26,33]. Nevertheless, broader validation across geographically and ethnically diverse populations is required before the current decision-support system could be recommended for routine clinical use.

The present findings should be interpreted in light of several limitations. The most important limitation is the simulated nature of the study. Although the dataset was based on complete digital records from patients who underwent bimaxillary orthognathic surgery, the AI recommendations did not influence actual treatment. The reported agreement, efficiency, and usability metrics therefore demonstrate technical and simulated clinical

performance rather than improved real-world patient outcomes. A prospective interventional trial is required to determine whether AI-assisted planning improves operative accuracy, postoperative function, facial aesthetics, stability, quality of life, or complication rates.

Second, expert consensus was used as the reference standard. Consensus among experienced surgeons is an appropriate comparator for treatment planning, but it does not represent an absolute biological truth. Different surgical strategies may produce similarly favorable outcomes, and experts may share institutional biases. High agreement with experts therefore indicates clinical plausibility rather than proof that the AI recommendation is objectively optimal. Future studies should compare alternative treatment recommendations against actual longitudinal outcomes and patient-reported satisfaction.

Third, the same general research program contributed to the development of several analytical components incorporated into the present platform [15–33]. Although this continuity allowed integration of previously validated segmentation, planning, and prediction modules, it may increase the risk of correlated assumptions and methodological dependence. Independent external validation by research groups not involved in system development will be necessary.

Fourth, the composite Clinical Decision Score was based on weighted objectives. The relative weighting of facial aesthetics, stability, symmetry, airway effects, treatment complexity, and predicted satisfaction inevitably reflects clinical judgments. Different patients and institutions may prioritize these outcomes differently. A patient with obstructive sleep apnea may assign greater importance to airway improvement, whereas another patient may prioritize facial aesthetics or avoidance of genioplasty. Future systems should therefore permit adjustment of objective weights according to patient preferences and clinical context.

Fifth, several factors relevant to treatment planning were incompletely represented. These include detailed temporomandibular joint function, muscle activity, soft tissue biomechanics, psychological expectations, socioeconomic considerations, and surgeon-specific technical preferences. Although imaging and cephalometric variables provide extensive anatomical information, not every clinically relevant factor can be quantified within current models. Integration of patient-reported goals, functional examinations, and biomechanical modeling may improve future recommendations.

Sixth, the explainability methods have inherent limitations. SHAP and Grad-CAM improve transparency but do not fully reveal the internal reasoning of deep neural networks. Their outputs may be unstable under certain conditions and can be influenced by correlated variables. Explainability should therefore be validated for consistency and clinical usefulness rather than assumed to provide definitive causal interpretation [40–43,48].

Seventh, the number of experts was limited to five experienced oral and maxillofacial surgeons. Although this panel provided a robust consensus reference, broader evaluation including orthodontists, trainees, and surgeons from different institutions would provide a more comprehensive assessment of clinical applicability. Future studies should also investigate whether AI support provides greater benefit to less experienced clinicians than to experts.

Despite these limitations, the study possesses several strengths. The multicenter design, consecutive patient cohort, complete digital datasets, standardized workflow, structured expert consensus, repeated model analysis, internal cross-validation, confidence estimation, and integration of explainable AI provide a comprehensive evaluation of the proposed decision-support system. The simultaneous assessment of agreement, quantitative planning accuracy, efficiency, reproducibility, usability, and interpretability extends beyond the performance metrics commonly reported in medical AI studies.

The clinical progression represented by the present research program is also noteworthy. Earlier work established automated detection and planning in maxillofacial trauma [15–26]. Subsequent investigations examined AI-assisted virtual planning, transfer accuracy, long-term stability, soft tissue prediction, relapse prediction, and multicenter validation in orthognathic surgery [27–33]. The current study integrates these capabilities within a treatment recommendation framework. The next stage should involve prospective clinical implementation with predefined safety thresholds and direct measurement of patient outcomes.

A future randomized clinical trial could compare conventional multidisciplinary planning with AI-assisted planning. Primary outcomes should include planning time, frequency of plan revisions, agreement between clinicians, intraoperative transfer accuracy, postoperative skeletal position, soft tissue outcomes, complications, relapse, patient satisfaction, and cost-effectiveness. Clinicians should remain blinded to AI recommendations during control planning, while independent adjudicators assess outcome quality. Long-term follow-up will be necessary to determine whether AI recommendations translate into improved stability rather than only improved agreement.

Further research should also investigate adaptive and interactive decision support. Instead of generating a single final recommendation, the system could allow the surgeon to adjust priorities and observe how the optimal plan changes. Increasing the importance of airway expansion, for example, might produce a different recommendation from prioritizing facial symmetry or minimum relapse risk. This interactive approach would preserve clinician control while exploiting the computational capacity of AI.

Integration with foundation models may further expand capability. Current systems typically use specialized models for segmentation, landmark detection, prediction, and recommendation. Multimodal foundation models could potentially interpret CBCT images, facial photographs, dental scans, clinical text, and treatment history within a unified architecture [34–38,44–46]. However, larger models also introduce challenges related to transparency, computational requirements, hallucination, privacy, and regulatory oversight. Their use in surgical decision-making should therefore proceed cautiously and require rigorous validation.

Intraoperative integration represents another important direction. AI-assisted navigation could compare actual skeletal positioning with the planned movements and provide real-time warnings when deviations exceed predefined thresholds. Intraoperative imaging could update the digital patient model and recalculate predicted soft tissue outcomes or stability. Robotic or navigated instrumentation might eventually use these data to improve execution, although the surgeon should retain final control.

Patient involvement should likewise become a central component of future decision support. Personalized planning should not be based exclusively on anatomical optimization. Patients should be able to review alternative simulations, express preferences, and understand the trade-offs between aesthetics, stability, airway effects, and procedural complexity. Explainable AI reports may be adapted into patient-facing visualizations that support shared decision-making without overstating the certainty of predicted outcomes.

In conclusion, the present prospective multicenter simulation study demonstrated that a multimodal artificial intelligence-assisted clinical decision-support system can generate personalized orthognathic treatment recommendations with almost perfect agreement to an experienced expert consensus. The system achieved high quantitative accuracy, excellent reproducibility, substantial workflow efficiency, clinically plausible explainability, and favorable surgeon acceptance. These findings indicate that artificial intelligence may evolve from an analytical and predictive technology into an integrated clinical partner capable of supporting complex surgical decisions.

The system should not be interpreted as an autonomous replacement for the surgeon. Its principal value lies in standardized multimodal analysis, rapid evaluation of alternative treatment scenarios, transparent communication of influential factors, estimation of uncertainty, and identification of cases requiring additional expert review. Prospective real-world validation is essential before routine clinical implementation. If confirmed in independent and interventional studies, AI-assisted clinical decision support may contribute to more precise, efficient, transparent, and patient-centered orthognathic surgery.

5. CONCLUSION

The present prospective multicenter simulation study demonstrated that a multimodal artificial intelligence-assisted clinical decision-support system can generate individualized treatment recommendations for bimaxillary orthognathic surgery with almost perfect agreement to an experienced expert consensus panel.

The proposed framework achieved high quantitative accuracy for translational and rotational skeletal movements, excellent reproducibility, substantial reductions in treatment-planning time, and favorable surgeon-reported usability.

The integration of automated image analysis, cephalometric assessment, virtual surgical scenario generation, postoperative outcome prediction, confidence estimation, and explainable artificial intelligence within a unified platform represents an important progression beyond conventional virtual surgical planning. Rather than functioning solely as a geometric simulation or prediction tool, the system supported the selection and prioritization of personalized surgical strategies while providing transparent information regarding the anatomical and clinical factors underlying each recommendation.

Artificial intelligence should not be regarded as a replacement for surgical expertise. Its principal clinical value lies in augmenting multidisciplinary decision-making, standardizing complex analyses, rapidly comparing alternative treatment strategies, and identifying cases that require additional expert review. Prospective clinical validation in independent real-world cohorts remains necessary before routine implementation can be recommended.

If confirmed in future interventional studies, AI-assisted clinical decision support may contribute to more precise, efficient, reproducible, transparent, and patient-centered orthognathic surgery.

6. ETHICS STATEMENT

This study was conducted in accordance with the Declaration of Helsinki and approved by the institutional ethics committee of Seeklinik Zurich, Specialized Clinic for Oral, Maxillofacial and Plastic Facial Surgery, Zurich, Switzerland (Approval No. SZ-OMFS-2020-014). Written informed consent was obtained from all participants.

7. CONFLICTS OF INTEREST

The authors declare no conflicts of interest related to this study.

8. FUNDING

No external funding was received for this study.

9. DATA AVAILABILITY STATEMENT

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

10. REFERENCES

[1] Farrell BB, Franco PB, Tucker MR. Virtual surgical planning in orthognathic surgery. Oral Maxillofac Surg Clin North Am. 2014;26(4):459–473.

- [2] Stokbro K, Aagaard E, Torkov P, Bell RB, Thygesen T. Virtual planning in orthognathic surgery. *Int J Oral Maxillofac Surg*. 2014;43(8):957–965.
- [3] Bobek S, Farrell B, Choi C, et al. Virtual surgical planning for orthognathic surgery using digital data transfer. *J Oral Maxillofac Surg*. 2015;73:1896–1903.
- [4] Jaisinghani S, Khechoyan D. Virtual surgical planning in orthognathic surgery. *Atlas Oral Maxillofac Surg Clin North Am*. 2017;25:69–82.
- [5] Alkhayer A, Piffkó J, Lippold C, Segatto E. Accuracy of virtual planning in orthognathic surgery: a systematic review. *Head Face Med*. 2020;16:34.
- [6] Chin SJ, Wilde F, Neuhaus M, et al. Accuracy of virtual surgical planning of orthognathic surgery with CAD/CAM-fabricated splints. *J Craniomaxillofac Surg*. 2017;45:1972–1980.
- [7] Barone M, Razionale AV, et al. The accuracy of jaw repositioning in bimaxillary orthognathic surgery using traditional and digital planning. *J Pers Med*. 2020;10:184.
- [8] Wong A, Cheung LK. Accuracy of maxillary repositioning surgery using CAD/CAM titanium surgical guides and fixation plates. *Clin Oral Investig*. 2021;25:2567–2575.
- [9] Tondin GM, et al. Evaluation of the accuracy of virtual planning in bimaxillary orthognathic surgery: a systematic review. *Int J Oral Maxillofac Surg*. 2022;51:114–123.
- [10] Stamm T, et al. In vivo accuracy of a new digital planning system in orthognathic surgery. *J Clin Med*. 2022;11:3112.
- [11] Shirota T, et al. CAD/CAM splint and surgical navigation allow accurate maxillary positioning. *Oral Maxillofac Surg*. 2019;23:45–52.
- [12] vMalenova Y, et al. Accuracy of maxillary positioning using computer-designed guides. *Clin Oral Investig*. 2023;27:6151–6161.
- [13] Trevisiol L, et al. Accuracy of virtual surgical planning in bimaxillary orthognathic surgery. *J Craniomaxillofac Surg*. 2023;51:502–510.
- [14] Li B, et al. Randomized clinical trial of patient-specific implants in orthognathic surgery. *J Craniomaxillofac Surg*. 2021;49:1024–1032.
- [15] Yildirim A, Hertach R, Yildirim V. Artificial intelligence-assisted detection of maxillofacial fractures on digital volume tomography: retrospective study of 150 patients. *J Med Dent*. 2026;2(1):44–52.
- [16] Yildirim A, Hertach R, Yildirim V. External multicenter validation of an artificial intelligence system for cone-beam CT-based detection of maxillofacial fractures: robustness across a tertiary facial trauma clinic and an independent maxillofacial practice. *J Med Dent*. 2026;2(1):70–81.
- [17] Yildirim A, Hertach R, Yildirim V. Artificial intelligence-assisted decision support in emergency maxillofacial trauma imaging: development and validation of a CBCT-based clinical decision algorithm. *J Med Dent*. 2026;2(1):82–92.
- [18] Yildirim A, Hertach R, Yildirim V. Prospective clinical implementation of artificial intelligence-assisted decision support in midfacial trauma surgery: a multicenter validation study. *J Med Dent*. 2026;2(1):93–99.
- [19] Yildirim A, Hertach R, Yildirim V. Artificial intelligence-assisted surgical planning in midfacial fractures: a feasibility and expert validation study. *J Med Dent*. 2026;2(1):100–108.
- [20] Yildirim A, Hertach R, Yildirim V. Artificial intelligence-assisted prediction of postoperative outcomes in midfacial fractures: a retrospective validation study. *J Med Dent*. 2026;2(1):109–117.

- [21] Yildirim A, Hertach R, Yildirim V. Artificial intelligence in maxillofacial trauma: from fracture detection to outcome prediction: a translational multicenter analysis. *J Med Dent*. 2026;2(1):118–125.
- [22] Yildirim A, Hertach R, Yildirim V. Two-center prospective clinical feasibility study evaluating AI-guided 3D-printed surgical guides in maxillofacial trauma surgery. *J Med Dent*. 2026;2(2):15–24.
- [23] Yildirim A, Hertach R, Yildirim V. Randomized controlled trial evaluating AI-guided 3D-printed surgical guides versus conventional surgery in maxillofacial trauma. *J Med Dent*. 2026;2(2):25–34.
- [24] Yildirim A, Hertach R, Yildirim V. Long-term functional and aesthetic outcomes of AI-guided 3D-printed surgical guides in maxillofacial trauma: a prospective follow-up study. *J Med Dent*. 2026;2(2):35–45.
- [25] Yildirim A, Hertach R, Yildirim V. Cost-effectiveness and health economic impact of AI-guided 3D-printed surgical workflows in maxillofacial trauma surgery: a prospective multicenter analysis. *J Med Dent*. 2026;2(2):46–58.
- [26] Yildirim A, Hertach R, Yildirim V. Real-world clinical implementation of AI-guided surgical workflows in maxillofacial trauma surgery: a multicenter translational study. *J Med Dent*. 2026;2(2):59–71.
- [27] Yildirim A, Hertach R, Yildirim V. AI-assisted virtual surgical planning and 3D-printed splint transfer in bimaxillary orthognathic surgery. *J Med Dent*. 2026;2(2):72–84.
- [28] Yildirim A, Hertach R, Yildirim V. Three-dimensional accuracy of AI-assisted virtual surgical planning in bimaxillary orthognathic surgery: a prospective comparative validation study. *J Med Dent*. 2026;2(2):85–98.
- [29] Yildirim A, Hertach R, Yildirim V. Randomized controlled trial comparing AI-assisted and conventional virtual surgical planning in bimaxillary orthognathic surgery. *J Med Dent*. 2026;2(2):99–110.
- [30] Yildirim A, Hertach R, Yildirim V. Long-term skeletal stability and patient-reported outcomes following AI-assisted virtual surgical planning in bimaxillary orthognathic surgery. *J Med Dent*. 2026;2(2):111–116.
- [31] Yildirim A, Hertach R, Yildirim V. AI-assisted soft tissue prediction and facial symmetry analysis following bimaxillary orthognathic surgery: a prospective three-dimensional clinical study. *J Med Dent*. 2026;2(2):117–126.
- [32] Yildirim A, Hertach R, Yildirim V. Artificial intelligence-assisted prediction of skeletal relapse and long-term stability following bimaxillary orthognathic surgery: a prospective three-dimensional analysis. *J Med Dent*. 2026;2(2):127–138.
- [33] Yildirim A, Hertach R, Yildirim V. Multicenter external validation of AI-assisted virtual surgical planning in bimaxillary orthognathic surgery: a comparative Swiss-German clinical study. *J Med Dent*. 2026;2(2):139–149.
- [34] Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259–265.
- [35] Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman BA, et al. UNETR: transformers for 3D medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022:574–584.
- [36] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203–211.
- [37] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment Anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023:4015–4026.

- [38] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations*. 2021.
- [39] Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med*. 2020;3:17.
- [40] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–4774.
- [41] Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?” Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:1135–1144.
- [42] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017:618–626.
- [43] Amann J, Blasimme A, Vayena E, Frey D, Madai VI; Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak*. 2020;20(1):310.
- [44] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56.
- [45] Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg*. 2018;268(1):70–76.
- [46] Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial intelligence and surgical decision-making. *JAMA Surg*. 2020;155(2):148–158.
- [47] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2019;9(4):e1312.
- [48] Corral-Acero J, Margara F, Marciniak M, Roderio C, Loncaric F, Feng Y, et al. The digital twin to enable the vision of precision cardiology. *Eur Heart J*. 2020;41(48):4556–4564.
- [49] Björnsson B, Borrebaeck C, Elander N, Gasslander T, Gawel DR, Gustafsson M, et al.; Swedish Digital Twin Consortium. Digital twins to personalize medicine. *Genome Med*. 2020;12(1):4.
-