Artificial intelligence (AI) for predicting the Aesthetic Component (AC) of the Index of Orthodontic Treatment Need (IOTN)

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Leah Stetzel, DDS

Graduate Program in Dentistry

The Ohio State University

2023

Thesis Committee

Ching-Chang Ko, DDS, MS, PhD, Advisor

Henry Fields, DDS, MS, MSD

Tai-Hsien Wu, PhD

Fernanda Schumacher, PhD

Stephen Richmond, BDS, DOrth RCS, MSD, FDS RCS, PHD, FHEA

Abstract

Introduction: One of the most widely used assessments of orthodontic treatment need is the Index of Orthodontic Treatment Need (IOTN). Multiple studies have verified the reliability and validity of the IOTN. The IOTN-Aesthetic Component (AC) defines esthetic impairment into ten levels; Level 1 represents the least treatment need, while Level 10 represents great need. However, the grading of IOTN is subjective. In this project, we propose the use of artificial intelligence (AI) to augment IOTN assessment which would allow for objective diagnoses, a reduced workload for orthodontists, at-home assessments of orthodontic treatment need, and potential utilization by third-party payers.

Objectives: The specific aim of this study was to collect a dataset of patients' oral images with the corresponding IOTN-AC classification and propose a deep-learning based algorithm that could identify the need for orthodontic treatment using intraoral photos.

Methods: 500 pre-treatment frontal intraoral photos with corresponding overjet values were collected. Each photo with overjet was graded by a gold standard IOTN rater. Intra-rater reliability was assessed. ResNet AI was trained using the verified intraoral images, overjet, and two different schemes (Scheme 1 and Scheme 2). The training data was annotated as 1-10 (representing IOTN-AC Level) in Scheme 1 and as 1-3 (representing "No need, borderline need, and great need" as described in the literature) in Scheme 2.

Both schemes were tested to predict ternary groups of “no need”, “borderline need”, or “great need”.

Furthermore, both schemes were tested to predict binary groups of IOTN <6 or IOTN ≥ 6 (a classification used by the National Health Service). In addition, we tested how the model would perform without an overjet value. Finally, our dataset was increased to n=564, and statistical analyses were re-run.

Results: Our gold standard rater had intra-rater reliability using weighted kappa of 0.84 (95% CI 0.76-0.93). Scheme 1 had an average of 62% sensitivity, 79% specificity, 68% accuracy, a positive predictive value (PPV) of 74%, and a negative predictive value (NPV) of 83% in predicting the ternary groups (“no need”, “borderline need”, or “great need”). Scheme 1 had 95% sensitivity, 52% specificity, 76% accuracy, a PPV of 72%, and a NPV of 88% in predicting the binary groups (IOTN <6 or IOTN ≥ 6). Scheme 2 had an average of 52% sensitivity, 78% specificity, 67% accuracy, a PPV of 82%, and a NPV of 84% for predicting the ternary groups. Scheme 2 had 77% sensitivity, 66% specificity, 72% accuracy, a PPV of 74%, and a NPV of 69% in predicting the binary group. When overjet was omitted, accuracy decreased by 1% in both ternary and binary predictions. When the dataset was supplemented, on average, the tests increased in accuracy by 2% for the binary predictions and by 3% in the ternary predictions.

Conclusion: We have developed a ResNet AI system that can automatically predict treatment need based on IOTN-AC reference standards. Results can presumably be improved with an increase in sample and training size.

Dedication

This thesis is dedicated to my incredible parents, Mark and Laura Stetzel. I will never be able to adequately thank you for the opportunities you have given me, the sacrifices you've made, and the great example you've set for me.

Acknowledgements

I wish to acknowledge:

- My loved ones —both family and friends—for their encouragement and continued support both near and far
- My late grandfather, my father, and my sister for providing an example of what it means to be a true dental professional
- Dr. Ching-Chang Ko and Dr. Henry Fields for their guidance and mentorship in not only my research, but also my entire residency at The Ohio State University
- Dr. Tai-Hsien Wu for his expertise, patience, and kindness in teaching me more about artificial intelligence than I thought possible
- Dr. Stephen Richmond for his collaboration and commitment to this project from across the world
- Dr. Fernanda Schumacher for teaching me about the nuances of statistical science
- The late Dr. Allen Firestone in helping me initiate this project, connect me with the right people, and push me to improve
- Dr. Momoko Karashima for her help with data collection
- My educators at Bishop Dwenger High School, Miami University, and Indiana University School of Dentistry who supported me in my academic and professional goals
- My co-residents for providing me memories and friendships that will last a lifetime

Vita

2013..... Bishop Dwenger High School, Fort Wayne, Indiana
2016.....B.A. Biology, Miami University, Oxford, Ohio
2020.....D.D.S., Indiana University, Indianapolis, Indiana
2023..... M.S. Orthodontics, The Ohio State University, Columbus, Ohio

Fields of Study

Major Field: Dentistry

Specialty: Orthodontics

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

ORTHODONTIC TREATMENT AND THE INDEX OF ORTHODONTIC TREATMENT NEED

In 2016, total orthodontic expenditures in the United States neared $20 billion.[1] Following diagnostic and restorative procedures, visits for orthodontic procedures are the third largest treatment category in dentistry.[2] The importance of a smile is widely accepted not only by society but also by scientific literature.

In a study by Shaw[3], it was found that children's dental features affect viewer's perception of their attractiveness and personal characteristics such as intelligence and aggressiveness. Similar results were confirmed by Papio et al.[4] in the adult population. Another study in 2011[5] found that ratings of attractiveness, intelligence, conscientiousness, agreeableness, and extraversion differed significantly depending on dental relationships or occlusion. Subjects with normal occlusion were rated the most positively in these categories. Because of this evidence, orthodontic treatment to improve esthetics and related social, intellectual, and integrity judgements is sought by patients and also recommended by orthodontists.

One of the most widely used assessments of orthodontic treatment need is The Index of Orthodontic Treatment Need (IOTN). Multiple studies have verified the reliability of the IOTN and its use on international populations.[6–8] The IOTN has two components: the Dental Health Component (DHC) and the Aesthetic Component (AC).

The DHC consists of a 5-point scale based on occlusal traits such as missing

teeth, crossbites, displacement of contact points, overjet, and overbite, where Grade 1 signifies “no treatment need” and Grade 5 signifies “very great treatment need”. This evaluation can be completed using physical or digital casts.[9]

The AC consists of a 10-point scale illustrated by a series of photographs that represent different levels of dental attractiveness (Figure 1).[9] In utilizing the IOTN-AC, a rating of 1-10 is assigned for *overall* dental attractiveness rather than particular similarities to the photographs. The final value should reflect treatment need on the grounds of esthetic impairment and, consequently, the psychosocial need for orthodontic treatment.[10] These photographs were evaluated by a group of lay judges and deemed to be equidistantly spaced between Grades 1-10.[11]

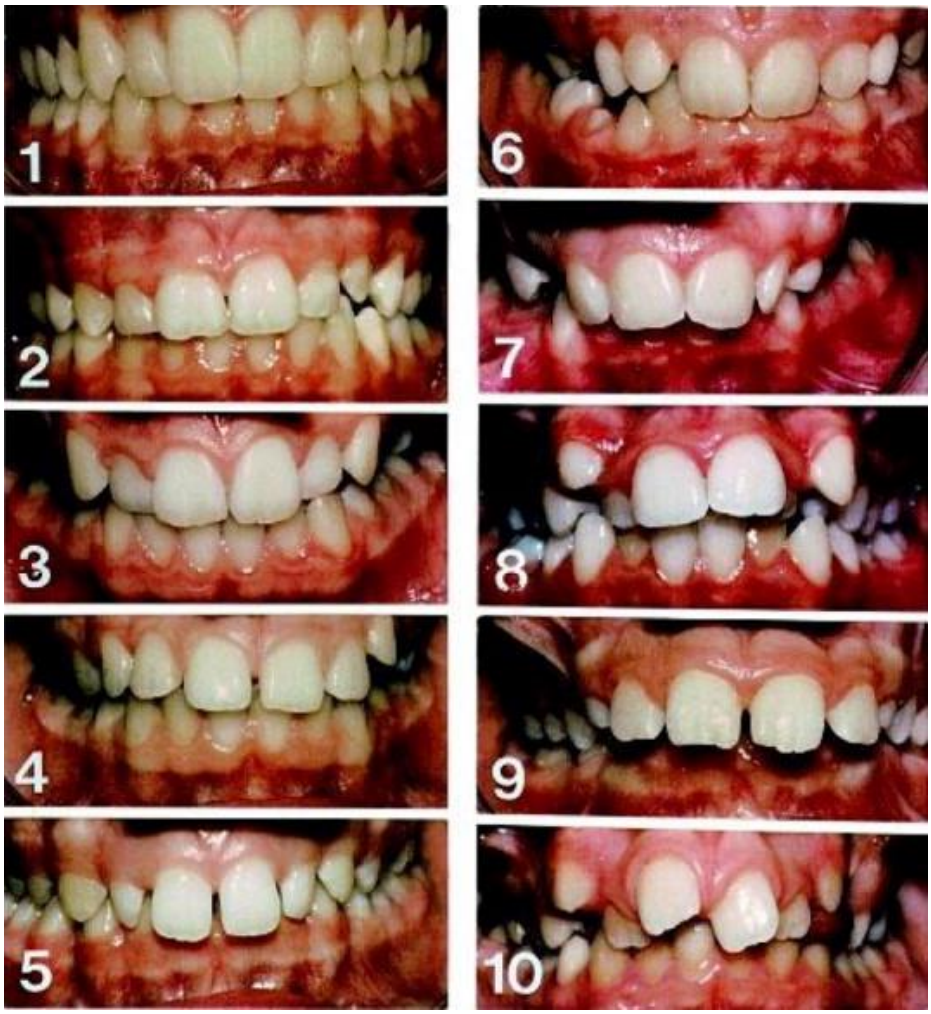

Figure 1. The Aesthetic Component of the Index of Orthodontic Treatment Need consists of 10-point scale illustrated by a series of photographs that represent different levels of dental attractiveness, where 1 represents an overall impression of little to no treatment need, and 10 represents an impression of great treatment need.[11]

With the use of a validation exercise, Richmond et al.[12] reported that the IOTN-AC grades could be partitioned into 3 treatment need subgroups: no need, borderline

need, and definite treatment need. IOTN-AC Grades 1-4, 5-7, and 8-10 were combined to signify no treatment need, borderline need, and great need, respectively, in this modified grouping.[13]

In the 2018 study by Papio et al.[4], the IOTN-AC was used to quantify how dental esthetics contribute to overall facial attractiveness. Patients with attractive, average, and unattractive faces, and with dental esthetics ranging from 1 to 10 according to the IOTN-AC, were rated with a lips together and lips apart pose. The differences provided a quantification of what esthetic dental alignment added to facial attractiveness. In females, only nearly ideal teeth (IOTN 1) can improve overall facial attractiveness. When dental esthetics were less than ideal (at IOTN 5 or more), dental impact was neutral or negative for all background facial attractiveness. On more attractive faces, low dental esthetics had a greater effect on overall attractiveness. This study also reported that men with an attractive face can camouflage dental deficits better than female counterparts.

Certain occlusal disharmonies, such as hypodontia or posterior crossbites, may have dental implications but are not considered esthetic effects in the IOTN, as they may not been seen from a frontal intraoral photo. Furthermore, using frontal intraoral photographs in the IOTN-AC rating limits overjet and incisor-to-lip evaluations.[14] These discrepancies can lead to disagreements between DHC and AC grades. In a recent study, it was demonstrated that only a moderate agreement between DHC and AC exists.[9] This level of agreement and the evaluation of only the frontal view of occlusion highlighted the subjective nature of the AC. In practical use, however, practitioners are calibrated for

IOTN using dental casts, so overjets and overbites are more precise than those portrayed in photographs. Additionally, when grading IOTN clinically, measurements for both DHC and AC are taken directly on the patient. Therefore, for some applications of the IOTN-AC, overjet is provided along with the frontal photograph. This accounts for inconsistencies in images and type of photoflash used.[15]

The IOTN is arguably most relevant in England and Wales, as it is currently used by the National Health Service (NHS) to determine whether children qualify for covered orthodontic treatment. Patient's with IOTN-DHC of 4 or 5 are eligible for NHS orthodontic treatment. However, the decision on treatment for borderline malocclusions, such as those with DHC of 3, is known to be difficult.[16,17] In 2006, a prioritization system was introduced so that these borderline cases (DHC=3) required an AC of Grade 6 or more in order to receive access to care with the NHS.[18] It is clear the AC evaluation impacts the ability of patients to receive care.

MACHINE LEARNING

Certain limitations exist in the conventional method of patient diagnosis. Information gathered from models, interviews, radiographs, and chair-side examinations are interpreted by clinicians with potential variations present at each step. Due to the individuality of the doctor's experience, these methods of data gathering and analysis provide large variations in decision outcomes and are largely empirical.

Today, artificial intelligence (AI), can be used to teach human learning processes to a machine. AI attempts to recognize human behavioral patterns in order to close

educational gaps and reduce variations intrinsic to human learning.[19]

Machine learning is a branch of AI. Traditional machine learning requires a large amount of data and is computationally expensive. Random forest and support vector machine (SVM) are popular machine learning methods. Deep learning, and specifically transfer learning, on the other hand, is more efficient and can be used with a smaller data set. Deep learning uses multiple layers of neural network algorithms to extract higher-level features from the data set. Transfer learning utilizes a pre-trained model to initiate a model for a new task. "Transfer learning refer(s) to the situation where what has been learned in one setting…is exploited to improve generalization in another setting."[20] Transfer learning is increasingly common and it is now rare to train a model from scratch. Common pre-trained models that exist and are successful at classifying images include ImageNet, AlexNet, Inception, and ResNet. These are examples of neural networks. This study took advantage of transfer learning and utilized the ResNet pre-trained model or neural network.

MACHINE LEARNING IN MEDICINE AND DENTISTRY

The use of AI has aided the medical field in diagnosis automation. Features of diabetic retinopathy, macular degeneration, and glaucoma can be identified with high sensitivity and specificity. In CT scans and chest x-rays used in respiratory medicine, pathology such as lung nodules or cancers can be identified by AI. Deep learning also has high diagnostic accuracy for breast cancer in multiple imaging modalities, such as mammogram, ultrasound, and digital breast tomosynthesis.[21]

Research on the use of AI in dentistry is encouraging for diagnosis of dental caries, identification of cephalometric landmarks, segmentation of maxillofacial cysts and tumors, diagnosis of root fractures, periapical pathology, and bone resorption.[22–24]

Orthodontics may be one of the earliest dental specialties to adapt AI into its practice.[19] Lateral cephalometric radiographs are routinely used by orthodontists to evaluate skeletal features, such as the relationship of the jaws to each other, the relationship of the jaws to the face, and the relationship of the jaws to the teeth. Lateral cephalometric films are also extremely important in orthognathic surgical planning.[25] When analyzing cephalometric films, a major task is landmark identification. Automating this task reduces the workload for orthodontists and surgeons, and has been identified as useful.[26] A systematic review and meta-analyses was conducted in 2021 to examine the accuracy of deep learning for detecting landmarks on cephalometric radiographs. From the 19 included studies, 80% of the landmarks were identified within 2mm, demonstrating that deep learning shows relatively high accuracy for detecting landmarks on lateral cephalometric radiographs.[27] This review even noted that deep learning performs similar to seasoned clinicians, and perhaps even better than inexperienced ones.[28,29]

The decision to extract or not extract teeth is another important part of orthodontic practice. Recent studies have explored the use of neural networks to predict 5 patterns of tooth extraction: non-extraction, all first premolars, all second premolars, maxillary first premolars only, or maxillary first premolars with mandibular second premolars. The AI

was able to predict extraction vs. non-extraction with 94% accuracy, and the extraction pattern with 84% accuracy.[30]

Analyzing facial characteristics is also an important aspect of orthodontics. In a Japanese population, Murata et al. [31] used an AI system to identify clinically useful facial traits (e.g., facial asymmetry, lip protrusion, vertical proportions, and profile shape). In this study, an experienced orthodontist evaluated 1000 lateral and frontal images of patients for certain facial traits. The average number of facial traits identified by the experienced orthodontist was 6.5 and ranged from 1-18 traits. 900 of the patient images were used to train the model, and 100 were used to test the model. In testing the AI, the AI was able to identify these facial traits with 95% accuracy, 39% sensitivity, and 97% specificity.

Finally, AI's ability to assess orthodontic treatment need has been explored. In another study by Murata et al.[32], AI was able to classify patients into 5 orthodontic treatment need categories with 45% accuracy. The categories ranged from 1 (no need for treatment) to 5 (need for treatment) and were based on intraoral images taken from five different angles.

PRESENT STUDY

In this project, we proposed the use of artificial intelligence (AI) to augment the IOTN-AC assessment which would allow for more objective diagnoses, a reduced workload for orthodontists, at-home patient assessments, and potential utilization by third-party payers. The purpose of this study was to collect a dataset of US patient's oral

images with the corresponding IOTN-AC classification and develop a deep-learning based algorithm that could identify the IOTN-AC using frontal intraoral photos.

CHAPTER 2: MATERIALS AND METHODS

This project was exempt from the Institutional Review Board review.

DATA COLLECTION

500 intraoral images were gathered from The Ohio State University's Graduate Orthodontic Clinic and the University of North Carolina's Graduate Orthodontic Clinic. The 500 intraoral images were gathered in a quota-sampling manner, such that they mirrored the US population according to race and overjet values established from the epidemiological literature (Table 1).[33–35]

Race	**Overjet Range (mm)**							**Total**
	≤ -3	**-2 to 0**	**1 to 2**	**3 to 4**	**5 to 6**	**7 to 10**	**>10**	
White	6	15	127	116	39	13	2	318
Hispanic	1	5	30	45	6	2	1	90
Black	1	5	23	26	7	2	1	65
Asian	0	3	9	12	2	1	0	27
								500

Table 1. Demographic data of the sample (n=500) that was representative of the US population based on race and overjet values.

The 500 intraoral images with a corresponding overjet range (Figure 2) were then sent to Dr. Stephen Richmond, who served as our gold-standard IOTN-AC rater. Our gold-standard rater assigned each photo an IOTN-AC grade of 1-10. After a two-week wash-out period, our gold standard rater re-graded 100 images in order to test for

reliability using kappa agreement. Dr. Richmond is known to have helped with the development and implementation of the IOTN and he leads the only accredited and validated calibration course for the IOTN in the UK.

Figure 2. Example of intraoral photo with corresponding overjet range (mm) in upper left.

The results of our initial data collection provided a representation of IOTN-AC grades in the US population seeking orthodontic care (Table 2).

IOTN	Count	Percentage
1	3	1%
2	30	6%
3	44	9%
4	70	14%
5	72	14%
6	94	19%
7	96	19%
8	72	14%
9	11	2%
10	8	2%

Table 2. IOTN-AC grades for our sample, which was selected in a manner to be representative of the IOTN-AC grades in the US population seeking orthodontic care.

DEEP LEARNING

We developed a deep neural network, called the IOTN network, which takes two inputs and has three modules. The inputs are a 2D frontal intraoral image and an overjet numeric value. The modules consist of: a Convolutional neural network (CNN) module, an overjet module, and an output module, corresponding to the two inputs (Figure 3).

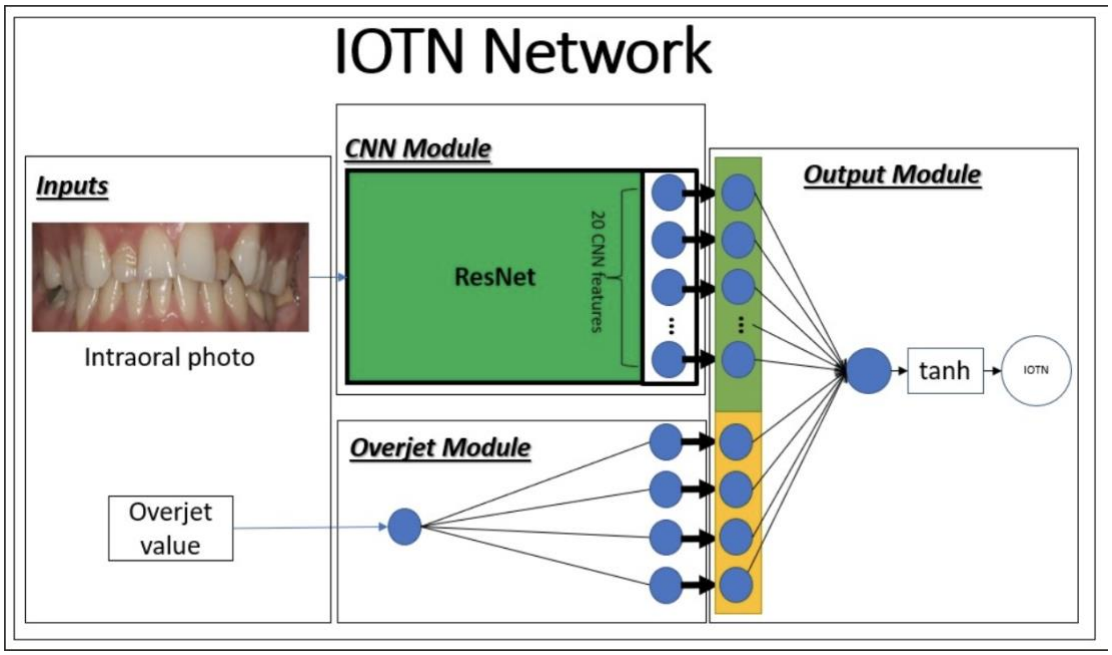

Figure 3. The IOTN network has 2 inputs, an overjet module, a CNN module, and an output module.

In the CNN module, we used Residual Network 34 (ResNet34), the most widely used neural network in both computer vision and medical imaging, to be the backbone to extract 20 hidden features.[36] In the overjet module, a two-layer fully connected network with the hyperbolic tangent activation function was used to learn 4 hidden features in an abstract domain from the overjet value. The 20 CNN features and 4 hidden overjet features were concatenated and feed into the final classification module to output the prediction. The output module is comprised of two fully connected layers followed by the hyperbolic tangent activation function.

It is worth noting that we consider this supervised task a regression problem

instead of a classification problem, since the IOTN grade implies the severity of the patient's oral conditions. For regression problems, the most commonly used activation function is hyperbolic tangent function and sigmoid function, whereas the softmax function is commonly used in classification problems. The IOTN Classification system uses integer numbers to represent treatment need, implying an ordinal relation. For example, patients with IOTN 1 (little to no need for treatment), will look more similar to those patients with IOTN 5-7 (borderline need for treatment) than they will to patients with IOTN 8-10 (great need for treatment). Therefore, the hyperbolic tangent activation function was adopted at the last layer in the output module instead of the softmax activation function. Furthermore, because the IOTN AC grades are equidistant from each other, it can be considered interval data.

IMPLEMENTATION

After our initial data collection of 500 photos, the IOTN network was *trained*, *validated*, and *tested* in a supervised learning manner. In machine learning, multiple models are often considered (or "trained") before a final model is chosen (or "validated"). The "validated" model is the one most optimized in terms of network parameters. The chosen, or "validated" model, is then "tested" with new, never-before-seen data in order to evaluate its performance and generalizability to unseen data.[37]

Of the 500 gathered, 360 images were used in the training phase, 40 images were used in the validation phase, and 100 were used in the testing phase. In the training phase, three inputs were given to the network: 1) an intraoral image, 2) an overjet value, and 3)

the gold standard (IOTN grade for that specific image determined by our gold standard orthodontic rater). Figure 4 shows a schematic of a training. The model took two inputs (intraoral image and overjet value) and was tasked with learning the output of IOTN. The predicted IOTN value was compared to the gold standard (third input) to calculate the discrepancy, which was back propagated to each layer of the network to update their parameters. To test the model, the AI was given unique, never before seen images with corresponding OJ values and was tasked with grading IOTN-AC.

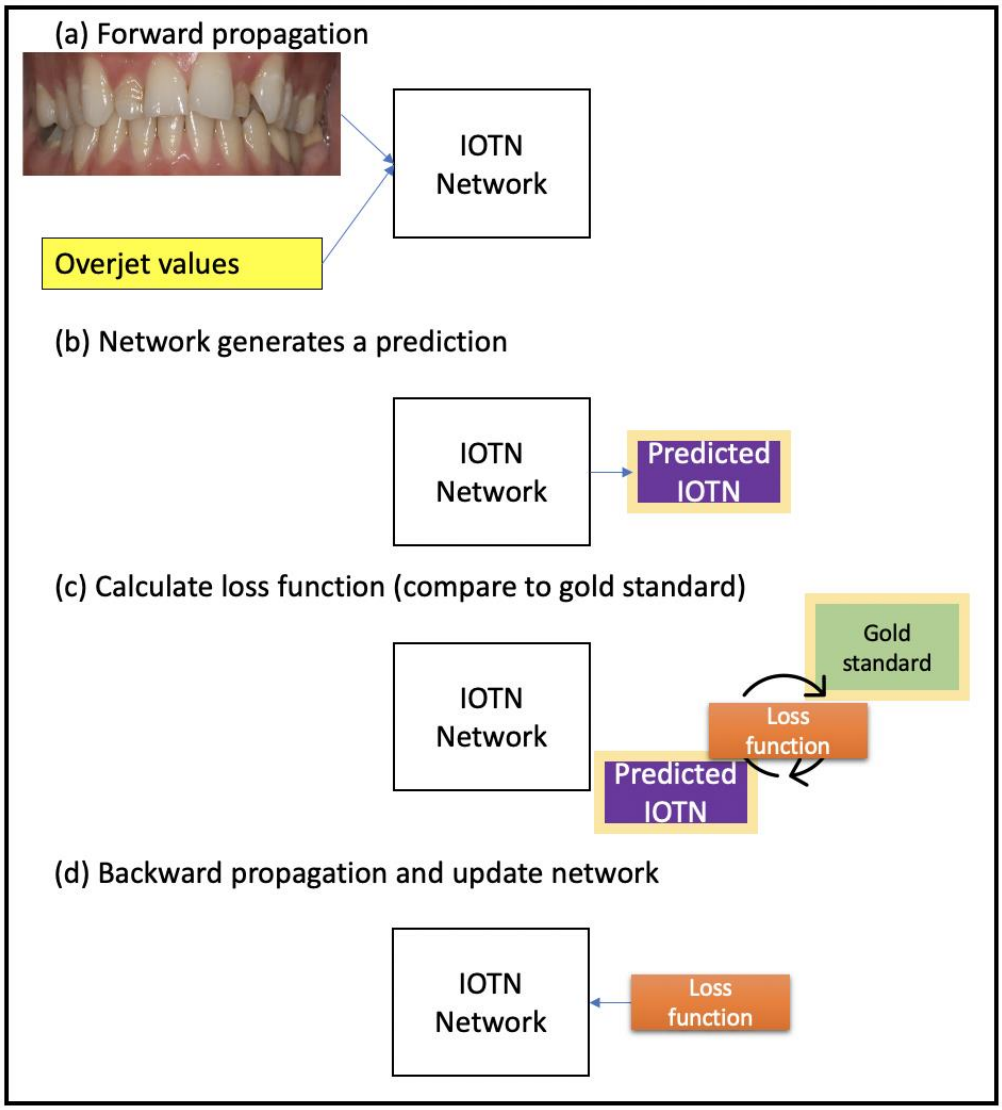

Figure 4. Schematic of a training where the model took the two inputs (i.e., intraoral image and overjet value) and was tasked with learning the output of IOTN. The predicted IOTN value was compared to the gold standard (third input) to calculate the discrepancy, which was back propagated to each layer of the network to update their parameters.

DATA AUGMENTATION AND TRANSFER LEARNING

To avoid the overfitting situation in our relatively small dataset, we adopted two techniques (data augmentation and transfer learning) to enhance the learning of visual representation. For data augmentation, we randomly applied different image filters on each image to "create" different images from the same source. The image filters used in this study include cropping and padding, sharpening, embossinig, Gaussian noise, Gaussian blur, contrast adjustment, and dropout (i.e., randomly removing some pixels). Each filter had random chances to be applied on the training images. By performing this heavy augmentation configuration, we expanded our training data to be 200 images for each grade, for a total 2000 images.

The second technique we applied was transfer learning, which indicates the process of applying previously acquired knowledge to new situations. This technique has been widely used in medical imaging studies since it is difficult to collect a large number of novel medical images. The pre-trained parameters of the ResNet34 previously trained by ImageNet (an open dataset containing 1,281,167 training natural images, 50,000 validation natural images, and 100,000 test natural images) for 1000 object classification were used in our CNN module. By doing this, our CNN module had an excellent initial ability to extract and recognize abstract features from intraoral photos since the network already could well recognize those natural images in ImageNet dataset. Then, we applied our augmented intraoral images to fine tune the CNN module in IOTN prediction. All the implementations were done by Pytorch, an open source deep learning library with Python programming language.[38] The data augmentation was carried out by imgaug, a library for

image augmentation in machine learning experiments. The pre-trained ResNet34 was downloaded from Pytorch. Some training configurations are provided in **Error! Reference source not found.**.

Optimizer	Adam
Learning rate	0.003
Epoch	200
Batch size	128
Loss function	Mean square error loss

Table 3. Details of the training configurations used in the CNN module.

OUTPUT AND TYPE OF PREDICTION

Prediction of 10 IOTN-AC levels- Scheme 0

As described in the implementation, the *training* model took the two inputs (i.e., intraoral image and overjet value) and was tasked with learning the output of IOTN. The predicted IOTN value was compared to the gold standard (third input to the training module) to calculate the discrepancy, which was back propagated to each layer of the network to update their parameters. In the *testing* phase of our first scheme, denoted as Scheme 0, the predicted IOTN grade 1-10 was compared to the gold standard of IOTN 1-10. The discrepancy between the predicted IOTN value and the gold standard allowed us to measure the performance of the AI system based on sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), and accuracy (ACC). Figure 5 illustrates Scheme 0.

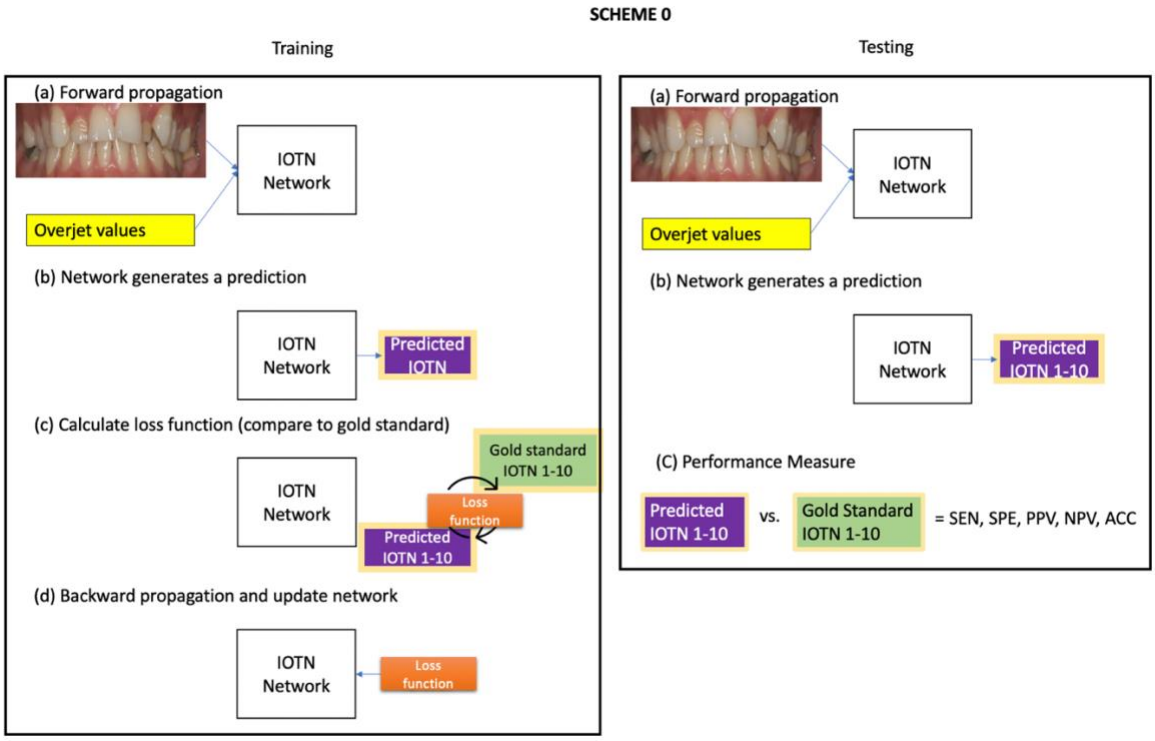

Figure 5. Scheme 0 training and testing, where the testing performance was based on the ability to predict IOTN 1-10.

Prediction of Simplified IOTN groups- Scheme 1 and Scheme 2

The task of predicting each IOTN-AC level proved challenging using only 500 intraoral photos. Therefore, instead of using 10 levels, the IOTN can be simplified to into binary or ternary classes. In the binary classification, group I corresponds to IOTN 1-5, where patients in the UK are denied treatment coverage, and group II corresponds to IOTN 6-10, where patients are granted orthodontic coverage with the NHS if their DHC is borderline (DHC=3).[18] The ternary classification is divided as follows: group I indicates no treatment need (corresponding to IOTN-AC 1-4), group II indicates

borderline need (corresponding to IOTN-AC 5-7), and group III indicates great need (corresponding to IOTN-AC 8-10). These cut off points are described in the literature.[7,13] To utilize the simplified IOTN classes, two additional implementation schemes were developed, called Scheme 1 and Scheme 2, in our AI system, described below.

In Scheme 1, the exact same training configuration as described in Scheme 0 was used. At the end of the testing phase, however, we added a procedure, called mapping, to simplify the IOTN-AC prediction and gold standard into binary (Figure *6*) or ternary (Figure 7) classes. In the binary classification, IOTN 1-5 was simplified to I, and IOTN 6-10 was simplified to II and in the ternary classification IOTN 1-4 was simplified to I, IOTN 5-7 was simplified to II, and IOTN 8-10 was simplified to III, as described above. The performance of the AI system was evaluated on the *simplified* classes instead of the original 10 IOTN-AC grades.

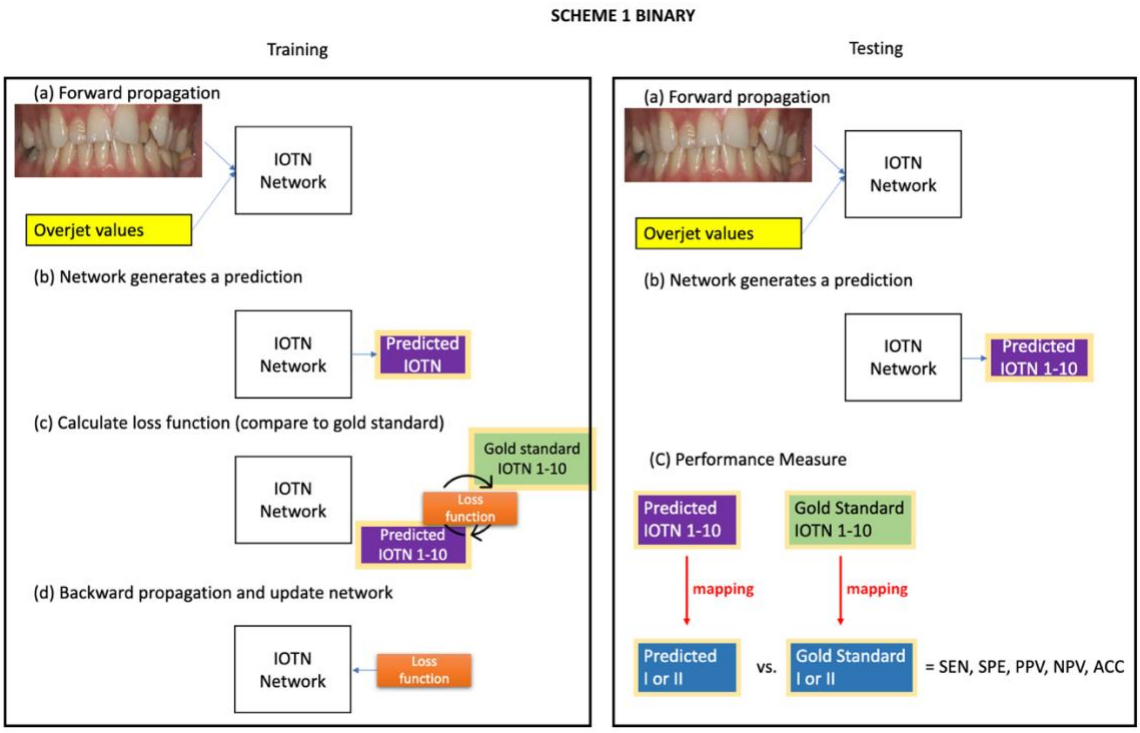

Figure 6. Scheme 1 with Binary Grouping, where the performance measure was based on simplified IOTN Classes of I (IOTN1-5) or II (IOTN 6-10).

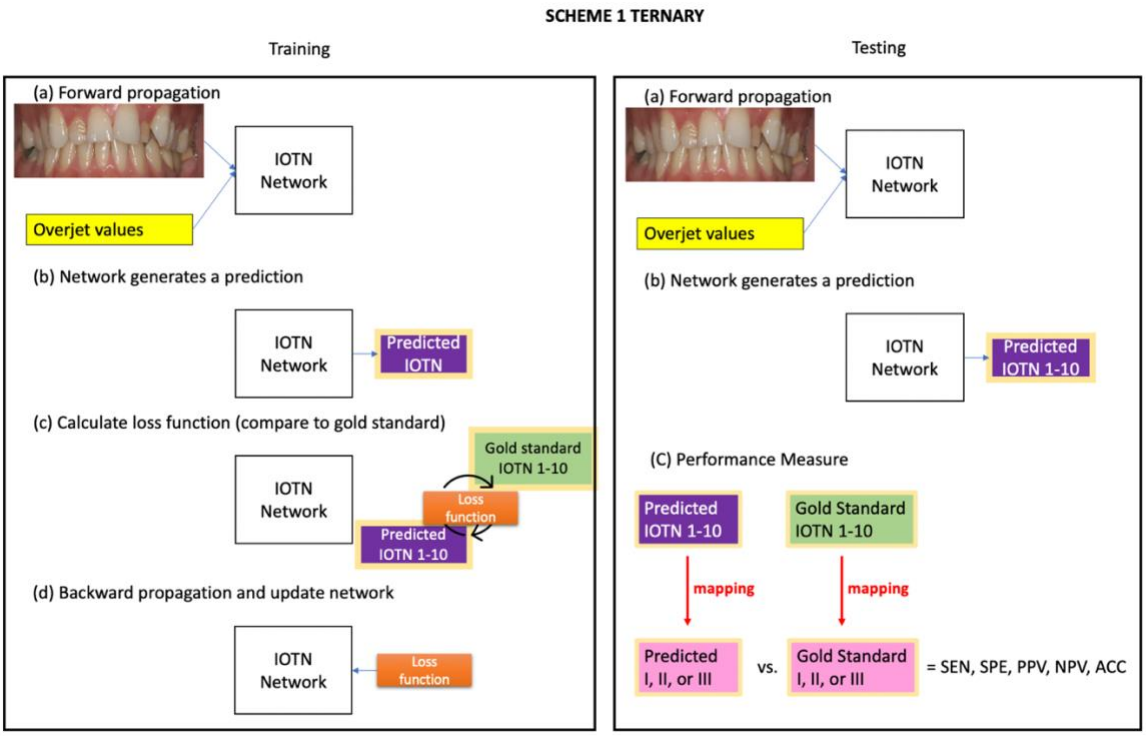

Figure 7. Scheme 1 with Ternary Grouping, where performance measure was based on simplified IOTN Classes of I (IOTN 1-4), II (IOTN 5-7), or III (IOTN 8-10).

In Scheme 2, a different *training* scheme was used where the input to the network was simplified. Instead of gold standard being IOTN 1-10, it was simplified to the binary or ternary treatment need categories within the network itself. Therefore, the output of IOTN network naturally became binary (Figure 8) or ternary classes (Figure 9).

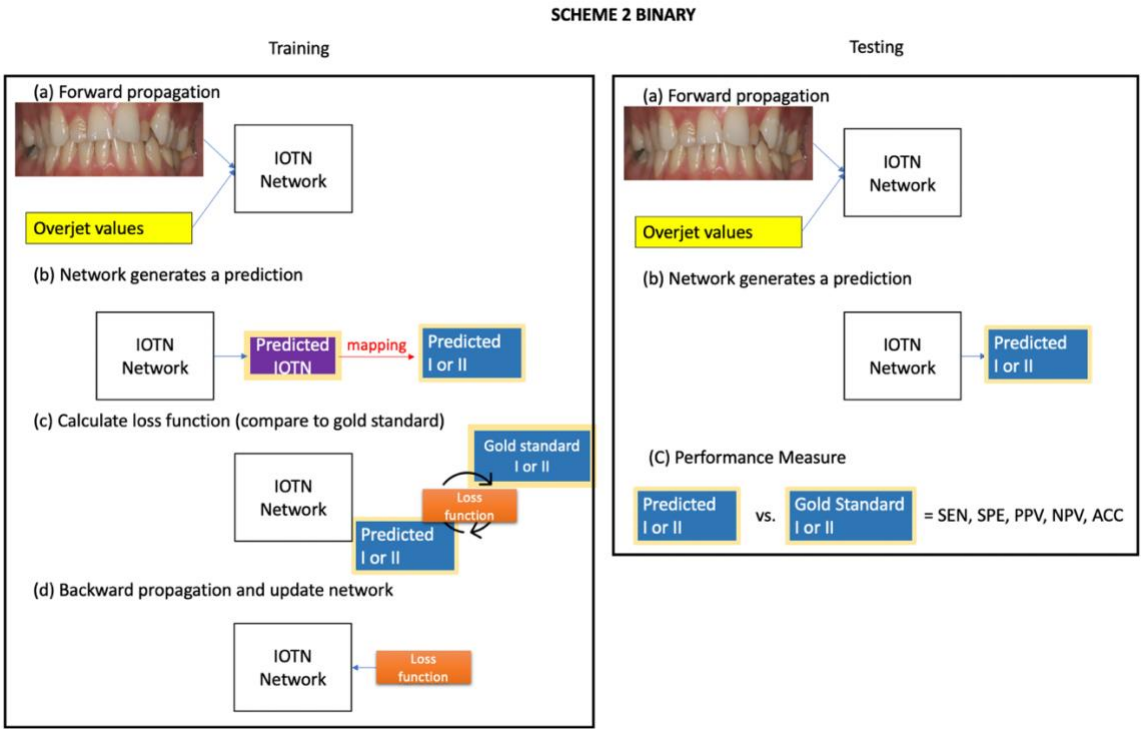

Figure 8. Scheme 2 with Binary Grouping, where training was based on simplified IOTN Classes of I (IOTN1-5) or II (IOTN 6-10).

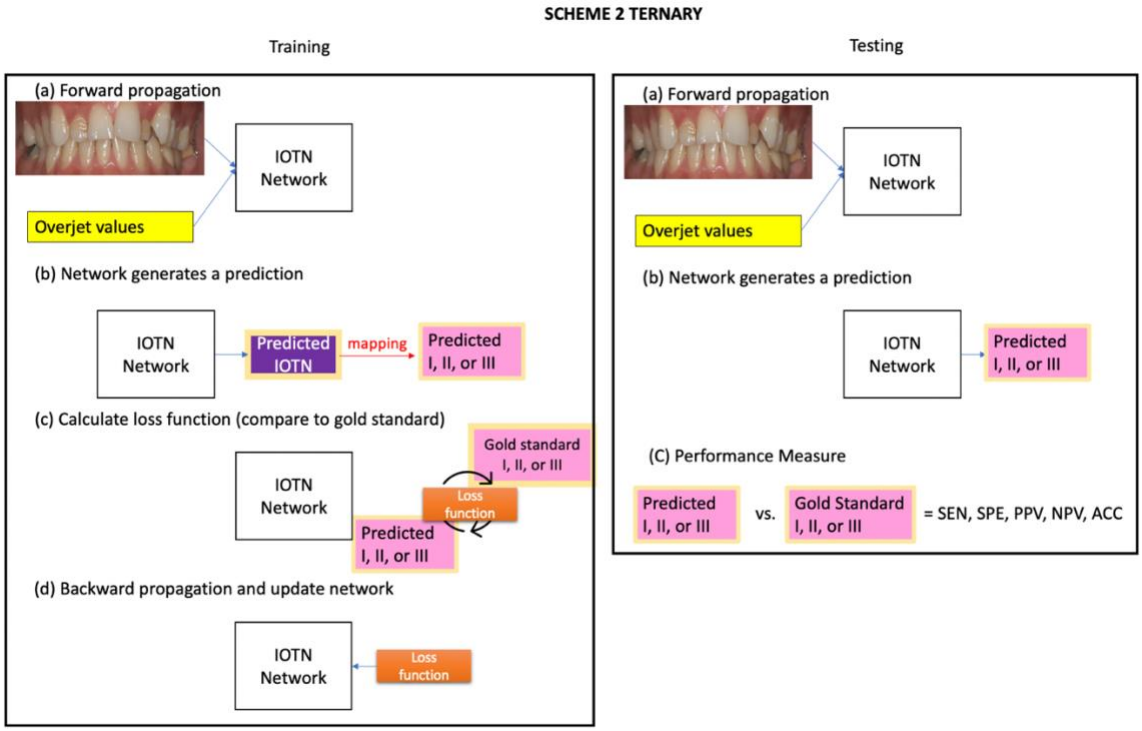

Figure 9. Scheme 2 with Ternary Grouping, where training was based on simplified IOTN Classes of I (IOTN 1-4), II (IOTN 5-7), or III (IOTN 8-10).

In summary, in Scheme 1 the gold standard was annotated as 1 to 10 (representing IOTN-AC Level) in the training phase, and in Scheme 2 the gold standard was simplified into binary or ternary classifications in the training phase. Both schemes performance measures were on the ability to categorize photos into the binary or ternary classifications. Scheme 1 and Scheme 2 were compared by measures of sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), and accuracy (ACC).

THE IOTN NETWORK VARIANT AND SUPPLEMENTED DATASET

In addition, we also developed an IOTN network variant which only takes

intraoral images without overjet input (i.e., removing the overjet module in the original IOTN network). SEN, SPE, PPV, NPV, and ACC were used to evaluate the performance.

To further test the influence of size of dataset on the overall AI system performance, 64 more intraoral images previously graded on the IOTN-AC were obtained from Dr. Richmond. The dataset was supplemented such that we had at least 20 images in each IOTN grade. See Table 2 for original dataset prior to supplementation. We trained and tested the IOTN network again using the Scheme 1 and calculated SEN, SPE, PPV, NPV, and ACC.

CHAPTER 3: MANUSCRIPT

Artificial intelligence (AI) for predicting the Aesthetic Component (AC) of the Index of Orthodontic Treatment Need (IOTN)

L. Stetzel[1], T. Wu[1], H. Fields[1], F. Schumacher[2], S. Richmond[3], C. Ko[1]

[1]Division of Orthodontics, The Ohio State University, 305 W. 12th Avenue, Columbus, OH, USA;
[2]Division of Biostatistics, The Ohio State University, 1841 Neil Avenue, Columbus, OH, USA;
[3]Department of Orthodontics, Cardiff University, Heath Park, Cardiff, CF14 4XY, Wales, UK

Abstract

Background: The Aesthetic Component (AC) of the Index of Orthodontic Treatment Need (IOTN) is internationally recognized as a reliable and valid method for assessing aesthetic treatment need.

Objective: To use artificial intelligence (AI) to automate the AC assessment.

Methods: 500 pre-treatment frontal intraoral photos with overjet values were collected. Each photo was graded by an experienced calibrated clinician. AI was trained using the intraoral images, overjet, and two different approaches. For Scheme 1, the training data were AC 1-10. For Scheme 2, the training data were either 2 groups: AC 1-5 and AC 6-10 or 3 groups: AC 1-4, AC 5-7 and AC 8-10. Sensitivity, specificity, positive predictive value, negative predictive value, and accuracy were measured for all approaches. The performance was tested without overjet values as input. Finally, the dataset was increased to n=564 so each AC grade had at least 20 samples in the training phase.

Results: The intra-rater reliability for the grader using kappa was 0.84 (95% CI 0.76-0.93). Scheme 1 had 95% sensitivity, 52% specificity, 76% accuracy, 72% PPV, and 88% NPV in predicting the binary groups. All other Schemes offered poor tradeoffs. Omitting overjet and dataset supplementation results were mixed depending upon perspective.

Limitations: The training data was limited by AC 1, 9, and 10's small cell counts of the proportional data.

Conclusions & Implications: We have developed deep learning based algorithms that can predict treatment need based on IOTN-AC reference standards. This may be a useful adjunct to clinical assessment of dental aesthetics.

Introduction

The National Health Service has been facing severe pressure to reduce costs due to consequences of the Covid-19 pandemic, chronic understaffing issues, and a deficit (1). Yet, a recent survey in May of 2021 of members of the British Orthodontic Society showed a marked increase in demand for orthodontic services (2). In 2020/21, the NHS expenditure on primary care orthodontic services totaled £306 million (3). It is increasingly important to distribute limited funds in a manner such that those in need of treatment are eligible for and obtain orthodontic services.

The importance of a smile is widely accepted not only by society but also by scientific literature. In a study by Shaw (4), it was found that children's dental features affect viewer's perception of their attractiveness and personal characteristics such as

intelligence and aggressiveness. Similar results were confirmed by Papio et al. (5) in the adult population. Another study in 2011 (6) found that ratings of attractiveness, intelligence, conscientiousness, agreeableness, and extraversion differed significantly depending on dental relationships or occlusion. Subjects with normal occlusion were rated the most positively in these categories. Because of this evidence, orthodontic treatment to improve esthetics and related social, intellectual, and integrity judgements is sought by patients and also recommended by orthodontists.

One of the most widely used assessments of orthodontic treatment need is the Index of Orthodontic Treatment Need (IOTN). Multiple studies have verified the reliability of the IOTN and its use on international populations (7–9). The IOTN has two components: the Dental Health Component (DHC) and the Aesthetic Component (AC).

The DHC consists of a 5-point scale based on occlusal traits such as missing teeth, crossbites, displacement of contact points, overjet, and overbite, where grade 1 signifies "no treatment need" and grade 5 signifies "very great treatment need". The AC consists of a 10-point scale illustrated by a series of photographs that represent different levels of dental attractiveness(10). In utilizing the IOTN-AC, a rating of 1-10 is assigned for *overall* dental attractiveness rather than particular similarities to the photographs. The final value should reflect treatment need on the grounds of aesthetic impairment and, consequently, the psychosocial need for orthodontic treatment(11).

With the use of a validation exercise, Richmond et al.(12) reported that the AC grades could be partitioned into 3 treatment need subgroups: no need, borderline need, and

definite treatment need. IOTN-AC grades 1-4, 5-7, and 8-10 were combined to signify no treatment need, borderline need, and great need, respectively, in this modified grouping(13).

The IOTN is currently used by the National Health Service (NHS) to determine whether children qualify for orthodontic treatment within the National Health Service. Patient's with IOTN-DHC of 4 or 5 are eligible for NHS orthodontic treatment. However, the decision on treatment for borderline malocclusions, such as those with DHC of 3, is known to be difficult(14,15). In 2006, a prioritization system was introduced so that these borderline cases (DHC=3) required an AC grade of 6 or more in order to receive eligibility for treatment with the NHS(16). It is clear the AC evaluation impacts the ability of patients to receive care.

With increasing demand for orthodontic care, reducing the workload of orthodontists and at-home patient assessments are appealing ideas. The use of artificial intelligence (AI) has aided both the medical and dental field in diagnosis automatization(17–20). Orthodontics may be one of the earliest dental specialties to adapt AI into its practice(21). A systematic review and meta-analysis were conducted in 2021 to examine the accuracy of deep learning (a branch of AI which utilizes neural networks) for detecting landmarks on cephalometric radiographs. From the 19 included studies, 80% of the landmarks were identified within 2mm, demonstrating that deep learning shows relatively high accuracy for detecting landmarks on lateral cephalometric radiographs(22). This review noted that deep learning performs similar to seasoned clinicians, and perhaps even better than

inexperienced ones(23,24).

AI's ability to assess orthodontic treatment need has been explored. In a study by Murata et al.(25), AI was able to classify patients into 5 orthodontic treatment need categories with 45% accuracy. The categories ranged from 1 (no need for treatment) to 5 (need for treatment) and were based on intraoral images taken from five different angles.

In this study, we proposed the use of artificial intelligence (AI) to augment the AC assessment which would allow for more objective assessments, a reduced workload for orthodontists, at-home patient assessments, and potential utilization by third-party payers. The purpose of this study was to collect a dataset of patients' oral images with the corresponding IOTN-AC classification and develop a deep-learning based AI algorithm that could identify the IOTN-AC using only a frontal intraoral photo and overjet range.

Methods

Data Collection

500 intraoral images were gathered in a quota-sampling manner, such that they mirrored the US population according to race and overjet values established from the epidemiological literature (26–28). The 500 intraoral images with a corresponding overjet range were assessed by an experienced calibrated examiner. Each photo was allocated an AC score. After a two-week wash-out period, 100 images were randomized to test for reliability using kappa. The 500 images then served as our gold standard.

Deep Learning

We developed a deep neural network, called the IOTN network, which takes two inputs

and has three modules. The inputs are a 2D frontal intraoral image and an overjet numeric value, which was the median value of the overjet range. The modules consist of a Convolutional neural network (CNN) module, an overjet module, and an output module, corresponding to the two inputs (Figure 1).

In the CNN module, we used Residual Network 34 (ResNet34), the most widely used neural network in both computer vision and medical imaging, to be the backbone to extract 20 hidden features (29). In the overjet module, a two-layer fully connected network with the hyperbolic tangent activation function was used to learn 4 hidden features in an abstract domain from the overjet value. The 20 CNN features and 4 hidden overjet features were concatenated and fed into the final classification module to output the prediction. The output module is comprised of two fully connected layers followed by the hyperbolic tangent activation function.

It is worth noting that we consider this supervised task a regression problem instead of a classification problem, since the IOTN-AC grade implies the severity of the patient's dental aesthetics. For regression problems, the most used activation functions are the hyperbolic tangent function and the sigmoid function. Whereas for classification problems, the softmax function is most used. The AC uses integer numbers to represent aesthetic treatment need, implying an ordinal relation. For example, patients with AC 1 (little to no need for treatment), look more similar to those patients with AC 5-7 (borderline need for treatment) than to patients with AC 8-10 (great need for treatment). Therefore, the hyperbolic tangent activation function was adopted as the last layer in the

output module instead of the softmax activation function. Furthermore, because the AC grades are equidistant from each other, it can be considered interval data.

Implementation

After our initial data collection of 500 photos, the IOTN network was trained, validated, and tested in a supervised learning manner. In machine learning, multiple models are often considered (or "trained") before a final model is chosen (or "validated"). The "validated" model is the one most optimized in terms of network parameters. The chosen, or "validated" model, is then "tested" with new, never-before-seen data in order to evaluate its performance and generalizability to unseen data (30).

Of the 500 gathered, 360 images were used in the training phase, 40 images were used in the validation phase, and 100 were used in the testing phase. In the training phase, three inputs were given to the network: 1) an intraoral image, 2) an overjet value, and 3) the gold standard (via the loss function, a measure of the difference between the gold standard and prediction). The discrepancy between the gold standard and the prediction was back propagated to each layer of the network to update their parameters. Figure 2 shows a schematic of how the IOTN network was trained.

To test the model, the AI was given 100 unique, new images with corresponding overjet value and was tasked with grading the AC. Figure 3 shows a schematic of the testing phase. The testing dataset mirrored the IOTN-AC distribution of our representative sample of 500. That is, the testing data had the same percentage of each AC grade as indicated in the initial data collection.

Data Augmentation and Transfer Learning

To avoid overfitting (when training results exceed those for novel data) in our relatively small dataset, we adopted two techniques: data augmentation and transfer learning. For data augmentation, we randomly applied different image filters on each image to "create" different images from the same source. The image filters used in this study include cropping and padding, sharpening, embossing, Gaussian noise, Gaussian blur, contrast adjustment, and dropout (i.e., randomly removing some pixels). Each filter had random chances to be applied on the training images. By performing this heavy augmentation configuration, we expanded our training data to be 200 images for each grade, for a total 2000 images. It is important to note that although each grade was augmented to have 200 images, the diversity of these grades was not equal.

The second technique we applied was transfer learning, which is the process of applying previously acquired knowledge to new situations. This technique has been widely used in medical imaging studies since it is difficult to collect a large number of novel medical images. The pre-trained parameters of the ResNet34 previously trained by ImageNet (an open dataset containing 1,281,167 training natural images, 50,000 validation natural images, and 100,000 test natural images) for 1000 object classification were used in our CNN module. By doing this, our CNN module had an excellent initial ability to extract and recognize abstract features from intraoral photos since the network already could recognize those natural images in ImageNet dataset. Then, we applied our augmented intraoral images to fine tune the CNN module in its ability to predict AC. All the

implementations were done by Pytorch, an open source deep learning library with Python programming language(31). The data augmentation was carried out by imgaug, a library for image augmentation in machine learning experiments. The pre-trained ResNet34 was downloaded from Pytorch.

Scheme 0

In the training phase of our first approach, denoted as Scheme 0, the gold standard was AC 1-10. In the testing phase of Scheme 0, the IOTN network predicted an AC grade 1-10 for each image.

Scheme 1

In the training phase of Scheme 1, the exact same training configuration as Scheme 0 was used, where the gold standard was AC 1-10. In the testing phase, however, we added a procedure, called mapping, at the end to simplify the AC prediction and gold standard into binary or ternary classes. In the binary classification, AC 1-5 was simplified to I, and AC 6-10 was simplified to II. In the ternary classification, AC 1-4 was simplified to I, AC 5-7 was simplified to II, and AC 8-10 was simplified to III.

Scheme 2

In the training phase of Scheme 2, the gold standard was simplified, into the binary and ternary groupings as described above. In Scheme 2, the IOTN network automatically predicted the simplified binary and ternary classifications, and mapping was unnecessary. A summary of the Schemes' trainings and tests can be found in Figure 4.

The IOTN Network Variant and Supplemented Dataset

In addition, we also developed an IOTN network variant which only takes the intraoral image as input (i.e., removes the overjet module in the original IOTN network).

To further test the influence of the size of the dataset on the overall AI system performance, 64 more intraoral images previously graded using the AC were obtained from the experienced calibrated examiner. The dataset was supplemented such that we had at least 20 images in each AC grade. 19 AC 1's, 20 AC 9's and 33 AC 10's were added to the extremes of the dataset. The IOTN network was trained and tested again using Scheme 1.

Statistical Analysis

All schemes' performances were measured by calculating sens, spec, PPV, NPV, and acc. For the binary predictions, an AC of 6-10 was considered a "positive" test and prediction, while AC 1-5 was considered a "negative" test and prediction.

For the ternary predictions, sens, spec, PPV, and NPV were calculated for each treatment need group I-III. For example, for the treatment need group III, a true positive was when the actual treatment need group was III and the predicted treatment need group was III. A false positive was when the actual treatment need group was either I or II and the predicted treatment need group was III. A true negative was when the actual treatment need group was either I or II and the predicted treatment need group was either I or II. Finally, a false negative was when the actual treatment need group was III and the predicted treatment need group was I or II.

Similarly, for the prediction of AC 1-10 (used in Scheme 0), sens, spec, PPV, and NPV

were calculated for each individual grade.

Results

The calibrated examiner demonstrated excellent intra-rater reliability in the identification of IOTN-AC grades 1-10 using kappa agreement, where the weighted kappa was 0.84 (95% CI 0.76 to 0.93).

The results of our initial data collection provided a representation of IOTN-AC grades in the US population. The most infrequent AC grade was 1, which represented 1% of our sample. AC 9 and 10 were also uncommon, and each represented 2% of our sample. AC 6 and 7 were the most frequent grades in our sample, each representing 19%. Complete AC distribution for our sample can be found in Table 1.

Predication of AC 1-10

For predicting AC 1-10, Scheme 0 had poor sensitivity, positive predictive value, and accuracy. When analyzing the performance of Scheme 0, 89% of errors (or absolute difference between gold standard and prediction when >0) were of either 1 or 2.

Prediction of AC 1-5 (I) and 6-10 (II) - Binary

For the binary predictions, Scheme 1 outperformed Scheme 2 in sensitivity, negative predictive value, and accuracy. Scheme 1 was able to identify those with AC 6-10 95% of the time. Scheme 2 had better specificity and PPV. The results of the binary predictions for Scheme 1 and Scheme 2 can be visualized in Figure 5.

Prediction of AC 1-4 (I), 5-7 (II), and 8-10 (III) - Ternary

For the ternary predictions, on average, Scheme 1 outperformed Scheme 2 in sensitivity,

specificity, and accuracy. Scheme 2 outperformed Scheme 1 with an average positive predictive value of 82% (8% greater than Scheme 1).

When analyzing the outcomes for each prediction group, it is evident that Scheme 1 misclassified actual “Borderline Need” subjects into both the “No Need” and “Great Need” category. Whereas Scheme 2 mis-predicted actual “Borderline Need” subjects into only the “No Need” category. Scheme 2 had substantially low sensitivity, and substantially high specificity and PPV. In this case, Scheme 2 mis-predicted all but 1 of the actual “Great Need” subjects into the “Borderline Need” group instead. Furthermore, there were no false positives for “Great Need” in Scheme 2. In both Scheme 1 and Scheme 2, the “Borderline Need” group had the highest sensitivity and the lowest specificity compared to both “No Need” and “Great Need” groups. These results can be visualized in Figure 6.

Predictions without overjet and with supplemented data

Without overjet, the model’s performance decreased in every metric on average for the ternary predictions. For the binary predictions, specificity and positive predicative value increased, while every other metric decreased.

When AC groups 1, 9, and 10 were supplemented with new images, the model’s performance increased in every metric on average for the ternary predictions. For the binary predictions, sensitivity and negative predicative value decreased, while every other metric increased. The results of the binary and ternary predictions with our supplemented data can be found in Figure 7.

Summary

The performance measures for all the schemes can be found in Table 2.

Discussion

In this study, we proposed the use of artificial intelligence to augment the AC assessment. We proposed multiple schemes of training and testing, and it is clear that results are variable depending upon how the AI model is trained and tested.

The experienced calibrated examiner had nearly perfect intra-rater reliability by weighted kappa, according to Cohen (32). This served as a strong underpinning for the study.

When originally attempting to classify the specific need categories of 1-10, our model (Scheme 0) proved inaccurate (acc=34%). However, when analyzing the discrepancies, or error, in this model, it was noted that 89% of errors were of only 1 or 2 grades, and a positive correlation was found (r=0.74). It is well-known that classification problems become more challenging as the number of classes increases, and a recent study suggests this increased complexity is due, at least in part to, the heterogeneity in decision boundaries.(33)

In order to improve our results, Scheme 1 and Scheme 2 were developed where the artificial intelligence was tasked to identify the broader treatment need categories (binary and ternary classifications). Emphasis was given to predicting these broader treatment need categories due to their practicality. The binary classification system is especially useful among those 18 years or younger enrolled in the NHS. If one is considered borderline in the DHC, the binary AC classification can determine if you are eligible for

NHS-funded treatment (AC 6-10) or if you will be ineligible (AC 1-5). The ternary classification is more descriptive where AC 1-4 indicates little to no treatment need, AC 5-7 indicates moderate treatment need, and AC 8-10 indicates great treatment need, but less useful in real application.

Certain metrics lead to the conclusion that Scheme 1 outperforms Scheme 2, and that Scheme 1 should be considered practically useful. The value of the outcomes really is one of perspective. If you are the payer, you do not want false positives, so high specificity and high PPV are critical. In fact, given the need to conserve funds for either the government or the administrator as net profit, you do not care about false negatives. From a patient or provider viewpoint, you want to minimize false negatives. So high sensitivity and high NPV are most important. You want all patients who need treatment to be appropriately allocated so it is of little concern if there are a few false positives, because all who qualify (and then some, maybe) will be funded.

It is important to note that certain third-party payers, such as the NHS, are funded by the public. According to the NHS Constitution for England Principles #2 and #6, "Access to NHS services is based on clinical need," and the NHS "is committed to providing the most effective, fair and sustainable use of finite resources"(34). Therefore, Scheme 1 with binary prediction could be considered for use by the NHS.

The ternary predictions may be clinically useful as they are more descriptive than the binary predictions. However, due to the poor sensitivity of the "Great Need" category, if this model was used to determine eligibility for care, many patients with true "Great

Need" for treatment would be mis-categorized as "Borderline Need". This may lead to an excess of appeals to third-party payers.

When analyzing at the binary grouping results (which is necessary when a patient has a DHC=3 in the NHS) of Scheme 1 vs. Scheme 2, Scheme 1 performed better overall. It would be desirable to have an automated system that can generate minimal false negatives (high sens), so that all of those needing treatment are identified. Furthermore, NPV in Scheme 1 was notably (20%) higher than in Scheme 2. This could assure patients, that if a negative result (IOTN-AC 1-5) is generated, likely (with 88% probability) a negative result was warranted.

Overall, the results of Scheme 1 were more promising than Scheme 2 when considering both binary and ternary predictions. Therefore, we decided to investigate how the Scheme 1 would perform without an overjet input. This would allow for less clinical error and less variation among practitioners. Accuracy decreased slightly (1%), but spec and PPV increased in the binary classification. From the perspective of patient and provider, removing the overjet input is currently not advisable, as sens and NPV decreased considerably without this input. In the ternary classification, all values decreased slightly, except spec, which stayed the same. This slight decrease in results may not be clinically significant. With an increase in sample size, it may be possible to classify IOTN-AC ranges without needing to measure overjet. This would allow for at-home patient assessments using a mobile device. More studies should be conducted.

We investigated how increasing sample or training size could impact our results. By

supplementing our dataset with an additional 64 images, we were able to ensure that each treatment need category had at least 20 images. In our original dataset (which represented the US population), AC Grades 1, 9, and 10 were substantially under-represented. When the data were supplemented, the prevalence was increased especially for AC 1, 9, and 10. Supplementing the dataset in this manner improved our results and it can be assumed that further increasing the sample size would further improve our results. Also, with an increase in prevalence, one would expect to see an increase in positive predictive value and a decrease in negative predictive value, and this was observed in the binary predictions. Again, this was not in the interest of the patients.

Additional ways to improve our results include improving the convolutional neural network.

Limitations of this study include a small sample size (AC 1, 9, and 10 had fewer representations). From a machine learning perspective, balanced training data is preferred. We have reason to believe that increasing the training dataset especially in these grades would improve the machine's ability overall to accurately predict the AC of the IOTN.

Conclusion

We have developed deep learning based algorithms that can predict dental aesthetic need based on IOTN-AC reference standards. Using the AC 1-10 scale input with binary testing was superior when compared to other AC categorizations and judged with a patient-centered public policy perspective.

References

1. An NHS under pressure [Internet]. BMA. 2022 [cited 2023 Feb 15]. Available from: https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/pressures/an-nhs-under-pressure
2. The 'Zoom Boom': BOS stats reveal a surge in demand for orthodontics during the pandemic [Internet]. British Orthodontic Society. 2021 [cited 2023 Feb 15]. Available from: https://www.bos.org.uk/ BOS-Homepage/News-Publications/Public-PatientsNews
3. Dental Services NHSBSA. England and Wales Orthodontic Contract Values 2020/21.
4. Shaw WC. The influence of children's dentofacial appearance on their social attractiveness as judged by peers and lay adults. Am J Orthod. 1981;79(4):399–415.
5. Papio MA, Fields HW, Beck FM, Firestone AR, Rosenstiel SF. The effect of dental and background facial attractiveness on facial attractiveness and perceived integrity and social and intellectual qualities. Am J Orthod Dentofac Orthop [Internet]. 2019;156(4):464-474.e1. Available from: https://doi.org/10.1016/j.ajodo.2018.10.021
6. Olsen JA, Inglehart MR. Malocclusions and perceptions of attractiveness, intelligence, and personality, and behavioral intentions. Am J Orthod Dentofac Orthop [Internet]. 2011;140(5):669–79. Available from: http://dx.doi.org/10.1016/j.ajodo.2011.02.025
7. Trivedi K, Shyagali TR, Doshi J, Rajpara Y. Reliability of Aesthetic component of IOTN in the assessment of subjective orthodontic treatment need. J Adv Oral Res. 2011;2(1):59–66.
8. Richmond S, Shaw WC, O'Brien KD, Buchanan IB, Stephens CD, Andrews M, et al. The relationship between the index of orthodontic treatment need and consensus opinion of a panel of 74 dentists. Br Dent J [Internet]. 1995;178(10):370–4. Available from: The relationship between the index of orthodontic treatment need and consensus opinion of a panel of 74 dentists

9. Younis J, Vig K, DJ R, RJ W. A validation study of three indexes of orthodontic treatment need in the United States. Community Dent Oral Epidemiol. 1997;25:358–62.
10. Borzabadi-Farahani A. An Overview of Selected Orthodontic Treatment Need Indices. In: Naretto S, editor. Principles in Contemporary Orthodontics [Internet]. Rijeka: InTech; 2011. Available from: http://www.intechopen.com/books/principles-in-contemporary-orthodontics/an-overview-of-selected- orthodontic-treatment-need-indices
11. Stenvik A, Espeland L, Linge BO, Linge L. Lay attitudes to dental appearance and need for orthodontic treatment. Eur J Orthod. 1997;19:271–7.
12. Richmond S, Shaw WC, O'Brien KD, Buchanan IB, Stephens CD, Andrews M, et al. The relationship between the index of orthodontic treatment need and consensus opinion of a panel of 74 dentists. Br Dent J [Internet]. 1995 May 20 [cited 2022 Oct 23];178(10):370–4. Available from: https://pubmed.ncbi.nlm.nih.gov/7779503/
13. Üçüncü N, Ertugay E. The use of the index of orthodontic treatment need (IOTN) in a school population and referred population. J Orthod. 2001;28(1):45–52.
14. Livas C, Delli K. Subjective and objective perception of orthodontic treatment need: A systematic review. Eur J Orthod. 2013;35(3):347–53.
15. Holmes A. The subjective need and demand for orthodontic treatment. Br J Orthod. 1992;19(4):287–97.
16. Needs Assessment for Orthodontic Services in London Needs Assessment for Orthodontic Services in London [Internet]. Public Health England. London; 2015. Available from: www.gov.uk/phe%0Awww.facebook.com/PublicHealthEngland
17. Shan T, Tay FR, Gu L. Application of Artificial Intelligence in Dentistry. J Dent Res. 2021;100(3):232–44.
18. Mohammad-Rahimi H, Nadimi M, Rohban MH, Shamsoddin E, Lee VY, Motamedian SR. Machine learning and orthodontics, current trends and the future opportunities: A scoping review. Am J Orthod Dentofac Orthop [Internet]. 2021;160(2):170-192.e4. Available from: http://dx.doi.org/10.1016/j.ajodo.2021.02.013
19. Mohammad-Rahimi H, Motamedian SR, Rohban MH, Krois J, Uribe SE,

Mahmoudinia E, et al. Deep learning for caries detection: A systematic review. J Dent [Internet]. 2022;122(January):104115. Available from: https://doi.org/10.1016/j.jdent.2022.104115

20. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. npj Digit Med. 2021;4.

21. Ko C-C, Tanikawa C, Wu T-H, Pastewait M, Bonebreak Jackson C, Kwon JJ, et al. EMBRACING NOVEL TECHNOLOGIES IN DENTISTRY AND ORTHODONTICS. In: Craniofacial Growth Series. 2020. p. 117–35.

22. Schwendicke F, Chaurasia A, Arsiwala L, Lee JH, Elhennawy K, Jost-Brinkmann PG, et al. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. Clin Oral Investig. 2021;25(7):4299–309.

23. Hwang HW, Park JH, Moon JH, Yu Y, Kim H, Her SB, et al. Automated identification of cephalometric landmarks: Part 2-Might it be better than human? Angle Orthod. 2020;90(1):69–76.

24. Muraev AA, Tsai P, Kibardin I, Oborotistov N, Shirayeva T, Ivanov S, et al. Frontal cephalometric landmarking: humans vs artificial neural networks. Int J Comput Dent [Internet]. 2020;23(2):139–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32555767

25. Murata S, Ishigaki K, Lee C, Tanikawa C, Date S, Yoshikawa T. Towards a Smart Dental Healthcare: An Automated Assessment of Orthodontic Treatment Need. Healthinfo. 2017;(c):35–9.

26. JONES N, MARKS R, RAMIREZ R, RÍOS-VARGAS M. 2020 Census Illuminates Racial and Ethnic Composition of the Country [Internet]. U.S. Census Bureau. 2020 [cited 2020 Dec 18]. Available from: https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html

27. Proffit WR, Fields HW, Moray LJ. Prevalence of malocclusion and orthodontic

treatment need in the United States: estimates from the NHANES III survey. Int J Adult Orthodon Orthognath Surg. 1998;13(2):97–106.

28. Alhammadi MS, Halboub E, Fayed MS, Labib A, El-Saaidi C. Global distribution of malocclusion traits: A systematic review. Dental Press J Orthod. 2018;23(6):e1–10.

29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016;2016-Decem:770–8.

30. Myrianthous G. Training vs Testing vs Validation Sets [Internet]. Towards Data Science. 2021 [cited 2022 Dec 15]. Available from: https://towardsdatascience.com/training-vs-testing-vs-validation-sets-a44bed52a0e1

31. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems [Internet]. Curran Associates, Inc.; 2019. Available from: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

32. McHugh ML. Lessons in biostatistics interrater reliability : the kappa statistic. Biochem Medica [Internet]. 2012;22(3):276–82. Available from: https://hrcak.srce.hr/89395

33. Moral P Del, Nowaczyk S, Pashami S. Why Is Multiclass Classification Hard? IEEE Access. 2022;10:80448–62.

34. The NHS Constitution [Internet]. [cited 2023 Feb 18]. Available from: https://www.gov.uk/government/publications/the-nhs-constitution-for-england/the-nhs-constitution-for-england

CHAPTER 4: RESULTS

Reliability

The gold standard IOTN grader demonstrated excellent intra-rater reliability in the identification of IOTN grades 1-10 using kappa agreement, where the weighted kappa was 0.84 (95% CI 0.76 to 0.93).

Prediction of 10 IOTN AC levels- Scheme 0

Scheme 0 yielded 34% accuracy (agreement between gold standard IOTN and predicted IOTN). Furthermore, 89% of our errors (or absolute difference between gold standard and prediction when >0) were errors of either 1 or 2. For example, if the gold standard was IOTN 2, but our AI predicted 4, the error (or difference between "actual" and "predicted") would be 2. Or if the gold standard was IOTN 3, but the AI predicted 4, the error would be 1. Table 4 shows the absolute difference between the gold standard and prediction with a count. Finally, there was a positive correlation between the gold standard IOTN and the predicted IOTN, represented in Figure 10 (r=0.74).

abs difference	count
0	34
1	41
2	18
3	4
4	3

Table 4. Absolute difference between gold standard IOTN and predicted IOTN using Scheme 1, along with count of the differences.

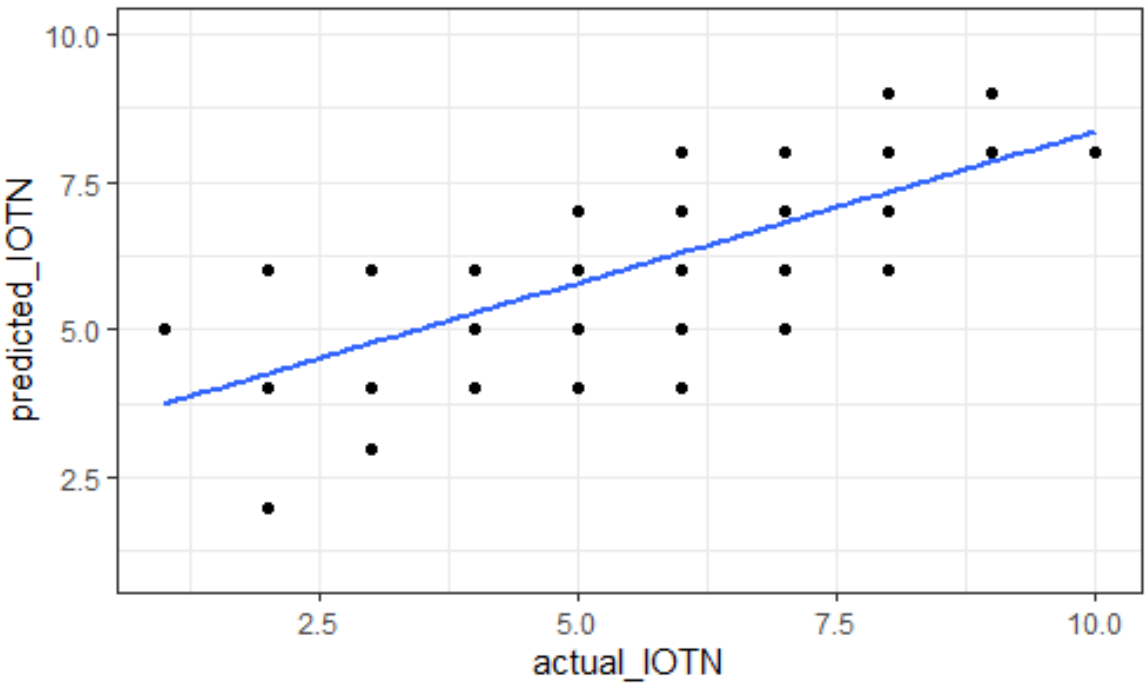

Figure 10. Scatterplot of Scheme 0.

Prediction of Simplified IOTN AC levels- Scheme 1

Scheme 1 yielded an accuracy of 76% and 68% for the binary and ternary

classifications, respectively.

The binary classification results are summarized in Table 5. The binary classification had a sensitivity (SEN) of 95% and a specificity (SPE) of 52%. The positive predictive value (PPV) was 72% and the negative predictive value (NPV) was 88%.

SEN	0.95
SPE	0.52
PPV	0.72
NPV	0.88

Table 5. Summary of Scheme 1 Binary Classification Results.

The ternary classification results are summarized in Table 6. The ternary classification's SEN ranged from 37% to 88%, while the SPE ranged from 46% to 97%. The PPV ranged from 64% to 85% and the NPV ranged from 78% to 92%. Furthermore, results of the ternary prediction for Scheme 1 can be visualized in Figure *11*

	No need (I)	**Borderline Need (II)**	**Great Need (III)**	**Average**
SEN	0.37	0.88	0.61	0.62
SPE	0.97	0.46	0.95	0.79
PPV	0.85	0.64	0.73	0.74
NPV	0.78	0.79	0.92	0.83

Table 6. Summary of Scheme 1 Ternary Classification Results.

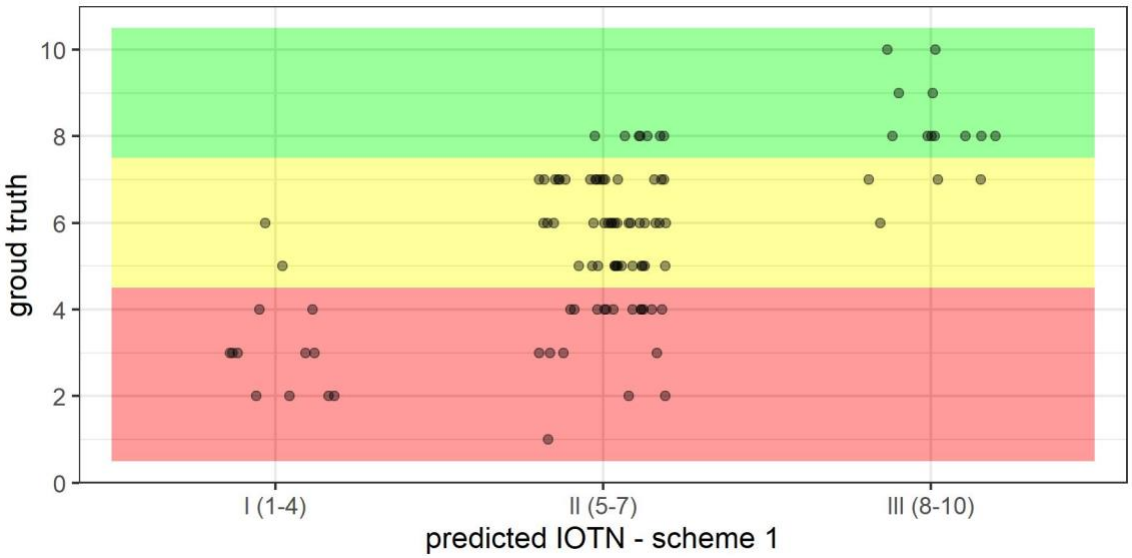

Figure 11. Scheme 1 with ternary grouping results.

A summary of Scheme 1 Binary vs. Ternary Classification can be found in Table 7, where the ternary classification values are averaged.

	Binary Classification	**Ternary Classification**
ACC	0.76	0.68
SEN	0.95	0.62
SPE	0.52	0.79
PPV	0.72	0.74
NPV	0.88	0.83

Table 7. Summary of Scheme 1 Binary vs. Ternary Classification Results.

Prediction of Simplified IOTN AC levels- Scheme 2

Scheme 2 yielded an accuracy of 72% and 67% for the binary and ternary classifications, respectively.

The binary classification results are summarized in Table 8. The binary classification had a SEN of 77% and a SPE of 66%. The PPV was 74% and the NPV was 69%.

SEN	0.77
SPE	0.66
PPV	0.74
NPV	0.69

Table 8. Summary of Scheme 2 Binary Classification Results.

The ternary classification statistical results are summarized in Table 9. The ternary classification's SEN ranged from 6% to 94%, while the SPE ranged from 38% to 100%. The PPV ranged from 62% to 100% and the NPV ranged from 83% to 86%. Furthermore, results of the ternary prediction for Scheme 2 can be visualized in Figure 12.

	No need (I)	Borderline Need (II)	Great Need (III)	Average
SEN	0.57	0.94	0.06	0.52
SPE	0.96	0.38	1	0.78
PPV	0.85	0.62	1	0.82
NPV	0.84	0.86	0.83	0.84

Table 9. Summary of Scheme 2 Ternary Classification Results

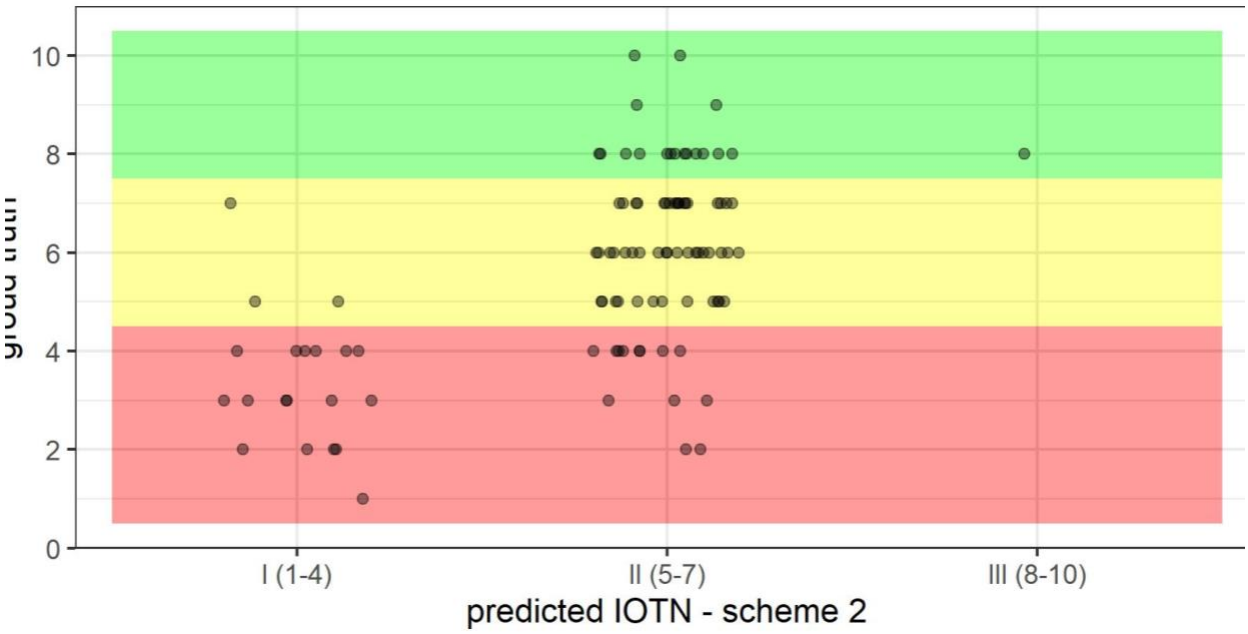

Figure 12. Scheme 2 with ternary grouping results.

A summary of Scheme 2 Binary vs. Ternary Classification can be found in Table 10, where the ternary classification values are averaged.

	Binary Classification	**Ternary Classification**
ACC	0.72	0.67
SEN	0.77	0.52
SPE	0.66	0.78
PPV	0.74	0.82
NPV	0.69	0.84

Table 10. Summary of Scheme 2 Binary vs. Ternary Classification Results.

Scheme 1 vs. Scheme 2

A summary of Scheme 1 vs. Scheme two results can be found in Table 11. Furthermore, confusion matrices can be found for the binary results in Table 12 and Table 13, and for the ternary results in Table 14 and Table 15.

	Scheme 1	**Scheme 2**
Binary Prediction	ACC: 0.76 SEN: 0.95 SPE: 0.52 PPV: 0.72 NPV: 0.88	ACC: 0.72 SEN: 0.77 SPE: 0.66 PPV: 0.74 NPV: 0.69
Ternary Prediction	ACC: 0.68 SEN: [0.37, 0.88, 0.61], μ=62 SPE: [0.97, 0.46, 0.95], μ=79 PPV: [0.85, 0.64, 0.73], μ=74 NPV: [0.78, 0.79, 0.92], μ=83	ACC: 0.67 SEN: [0.57, 0.94, 0.06], μ=52 SPE: [0.96, 0.38, 1.00], μ=78 PPV: [0.85, 0.62, 1.00], μ=82 NPV: [0.84, 0.86, 0.83], μ=84

Table 11. Summary of Scheme 1 vs. Scheme 2 Results.

	Predicted Positive	**Predicted Negative**
Actually Positive	53	3
Actually Negative	21	23

Table 12. Confusion Matrix for Scheme 1 Binary.

	Predicted Positive	**Predicted Negative**
Actually Positive	43	13
Actually Negative	15	29

Table 13. Confusion Matrix for Scheme 2 Binary.

	Predicted "No Need (I)"	**Predicted "Borderline Need (II)"**	**Predicted "Great Need (III)"**
Actual "No Need (I)"	11	19	0
Actual "Borderline Need (II)"	2	46	4
Actual "Great Need (III)"	0	7	11

Table 14. Confusion matrix for Scheme 1 Ternary.

	Predicted "No Need (I)"	**Predicted "Borderline Need (II)"**	**Predicted "Great Need (III)"**
Actual "No Need (I)"	17	13	0
Actual "Borderline Need (II)"	3	49	0
Actual "Great Need (III)"	0	17	1

Table 15. Confusion matrix for Scheme 2 Ternary.

Scheme 1 with No Overjet Values

When Scheme 1 was trained with no overjet input (only intraoral photo as input), its's accuracy decreased by 1% in the binary classification and decreased by 1% in the ternary classification. In the binary classification, SEN decreased by 25%, SPE increased by 30%, PPV increased by 11%, and NPV decreased by 20%.

Compared to the Scheme I ternary classification with overjet input, without overjet input, on average, the SEN decreased by 4% and the SPE stayed the same. In the no need category, SEN increased by 16% and SPE decreased by 3%. In the borderline need category, SEN decreased by 1% and SPE increased by 2%. In the great need category, SEN decreased by 28% and SPE stayed the same. On average, PPV decreased

by 6% and NPV decreased by 1%. In the no need category, PPV decreased by 5% and NPV increased by 5%. In the borderline need category, PPV stayed the same and NPV decreased by 2%. In the great need category, PPV decreased by 13% and NPV decreased by 5%.

A summary of Scheme 1 with and without overjet input can be found in Table 16.

	Scheme 1	**Scheme 1 w/out Overjet**
Binary Prediction	ACC: 0.76 SEN: 0.95 SPE: 0.52 PPV: 0.72 NPV: 0.88	ACC: 0.75 SEN: 0.70 SPE: 0.82 PPV: 0.83 NPV: 0.68
Ternary Prediction	ACC: 0.68 SEN: [0.37, 0.88, 0.61], μ=62 SPE: [0.97, 0.46, 0.95], μ=79 PPV: [0.85, 0.64, 0.73], μ=74 NPV: [0.78, 0.79, 0.92], μ=83	ACC: 0.67 SEN: [0.53, 0.87, 0.33], μ=58 SPE: [0.94, 0.48, 0.95], μ=79 PPV: [0.8, 0.64, 0.6], μ=68 NPV: [0.83, 0.77, 0.87], μ=82

Table 16. Summary of Scheme 1 with and without overjet input results.

Scheme 1 with Supplemented Dataset

With the additional 64 images, Scheme 1's accuracy improved by 2% in the binary classification and improved by 3% in the ternary classification.

In the Scheme I binary classification with supplemented data, sensitivity decreased by 16%, specificity increased by 25%, PPV increased by 9%, and NPV decreased by 14%.

For the ternary classification, on average, the SEN increased by 4% and the SPE increased by 3%. In the no need category, SEN increased by 16% and SPE decreased by 6%. In the borderline need category, SEN decreased by 3% and SPE increased by 10%. In the great need category, SEN was not changed and SPE increased by 3%. Furthermore, on average, PPV increased by 1% and NPV increased by 1%. In the no need category, PPV decreased by 12% and NPV increased by 4% with the supplemented data. In the borderline need category, PPV increased by 4% and NPV decreased by 2%. In the great need category, PPV increased by 12% and NPV stayed the same.

A summary of Scheme 1 vs. Scheme 1 supplemented can be found in Table 17. Additionally, a confusion matrix of Scheme 1 Supplemented Binary and Scheme 1 Supplemented Ternary can be found below in Table 18 and Table 19, respectively.

	Scheme 1	**Scheme 1 Supplemented**
Binary Prediction	ACC: 0.76 SEN: 0.95 SPE: 0.52 PPV: 0.72 NPV: 0.88	ACC: 0.78 SEN: 0.79 SPE: 0.77 PPV: 0.81 NPV: 0.74
Ternary Prediction	ACC: 0.68 SEN: [0.37, 0.88, 0.61], μ=62 SPE: [0.97, 0.46, 0.95], μ=79 PPV: [0.85, 0.64, 0.73], μ=74 NPV: [0.78, 0.79, 0.92], μ=83	ACC: 0.71 SEN: [0.53, 0.85, 0.61], μ=66 SPE: [0.91, 0.56, 0.98], μ=82 PPV: [0.73, 0.68, 0.85], μ=75 NPV: [0.82, 0.77, 0.92], μ=84

Table 17. Scheme 1 vs. Scheme 1 Supplemented Results

	Predicted Positive	**Predicted Negative**
Actually Positive	44	12
Actually Negative	10	34

Table 18. Confusion Matrix for Scheme 1 Supplemented Binary.

	Predicted "No Need (I)"	**Predicted "Borderline Need (II)"**	**Predicted "Great Need (III)"**
Actual "No Need (I)"	16	14	0
Actual "Borderline Need (II)"	6	44	2
Actual "Great Need (III)"	0	7	11

Table 19. Confusion Matrix for Scheme 1 Supplemented Ternary.

CHAPTER 5: DISCUSSION

Esthetic impairment can have profound social implications, and dental attractiveness significantly affects overall attractiveness. Orthodontic treatment is aimed at improving not only function, but also esthetics of the teeth and smile. Many individuals seek orthodontic care throughout the world.

The NHS in England and Wales utilizes the IOTN in order to determine who is eligible to receive orthodontic coverage. When the IOTN DHC is borderline, the AC becomes an important aspect in this decision. Here we proposed the automation of the IOTN-AC assessment, which would provide a more objective evaluation and minimize inevitable variation among practitioners.

In this study, emphasis was given to predicting broader treatment need categories. This included a binary classification (IOTN 1-5 or IOTN 6-10) and a ternary classification (IOTN 1-4, IOTN 5-7, and IOTN 8-10). The binary classification system is especially useful among those 18 years or younger enrolled in the NHS. If one is considered borderline in the DHC, the binary IOTN-AC classification can determine if you will receive coverage (IOTN 6-10) or if you will be denied coverage (IOTN 1-5). The ternary classification is more descriptive where IOTN 1-4 indicates little to no treatment need, IOTN 5-7 indicates moderate treatment need, and IOTN 8-10 indicates great treatment need.

When originally attempting to classify the specific need categories of IOTN 1-10, our model (Scheme 0) proved inaccurate (ACC=34%). However, when analyzing the

discrepancies, or error, in this model, it was noted that 89% of errors were of only 1 or 2 grades, and a positive correlation was found (r=0.74). It is well-known that classification problems become more challenging as the number of classes increases, and a recent study suggests this increased complexity is due, at least in part to, the heterogeneity in decision boundaries.[39]

In order to improve our results, Scheme 1 and Scheme 2 were developed where the artificial intelligence was tasked in identifying the broader treatment need categories (binary and ternary classifications). This strategy did notably improve our ACC, SEN, and PPV. Our SPE and NPV decreased with the simplification of classes.

In scheme 1, in training, the gold standard was annotated as 1 to 10 (representing specific IOTN-AC Grade), and in Scheme 2's training the gold standard was simplified into the broader binary or ternary classifications. Scheme 1, in general, had better results than Scheme 2. This can be rationalized by imagining the science of human learning. Scheme 1 is analogous to studying a textbook or novel and Scheme 2 is analogous to studying a summary or CliffsNotes. Both Schemes are given the same test, so it makes sense that Scheme 1 would perform better.

SCHEME 1 VS SCHEME 2

When analyzing at the Binary grouping results (which is necessary when a patient has a DHC=3 in the NHS) of Scheme 1 vs. Scheme 2, Scheme 1 performed better overall. It would be desirable to have an automated system that can generate minimal false negatives (high SEN), so that all of those needing treatment are captured. Furthermore,

NPV in Scheme 1 is notably (20%) higher than in Scheme 2. This can assure patients, that if you get a negative result (IOTN 1-4), you likely (with 88% probability) are truly IOTN 1-4.

Only SPE and PPV are higher in Scheme 2. From the perspective of the NHS or other third-party payers, it would be more costly to generate false positives. Therefore, Scheme 2 may be preferable to Scheme 1 from a cost standpoint, as Scheme 2 has less false negatives (higher SPE) than Scheme 1. Furthermore, we can be more confident that a positive test in Scheme 2 is truly someone that needs treatment (higher PPV).

When analyzing the Ternary Predictions of Scheme 1 vs. Scheme 2, our focus resides in the great need category, as care for these patients is most clinically relevant. The accuracy of Scheme 1 vs. Scheme 2's Ternary results are comparable. The sensitivity of Scheme 2, however, is poor and notably lower than that of Scheme 1. Scheme 2 was not able to identify true "great need", and instead often categorized these patients into the "borderline need" category instead. The PPV is then, justifiably, 100%. This means that the AI is only going to classify patients as "great need" if it is very confident that the patient is indeed "great need". Scheme 2 had a high specificity, meaning those who were true "borderline need" or true "no need" never fell into the great need category.

While SPE and PPV were lower in Scheme 1 compared to Scheme 2, Scheme 1's SPE and PPV were still promising (95% and 73%, respectively). Scheme 1 had better SEN and NPV than Scheme 2.

Overall, the results of Scheme 1 were more promising than Scheme 2 when considering both Binary and Ternary predictions. Therefore, we decided to investigate how the Scheme 1 would perform without an overjet input. This would allow for less clinical error and less variation among practitioners. Accuracy decreased slightly (1%), but SPE and PPV increased in the binary classification. In the ternary classification, all values decreased slightly, except SPE, which stayed the same. This slight decrease in results may not be clinically significant. With an increase in sample size, it may be possible to classify IOTN ranges without needing to measure overjet. This would allow for at-home patient assessments. More studies should be conducted.

We also investigated how increasing sample or training size could impact our results. By supplementing our dataset with an additional 64 images, we were able to ensure that each treatment need category had at least 20 images. In our original dataset (which represented the US population), IOTN Grades 1, 9, and 10 were significantly under-represented. Supplementing the dataset in this manner improved our results and it can be assumed that further increasing the sample size would further improve our results.

Additional ways to improve our results include improving the CNN or using a different training technique.

Limitations of this study include a small sample size. We have reason to believe that increasing the training dataset would improve the machine's ability to accurately predict IOTN-AC. Another limitation of the study is that only one gold-standard orthodontic rater was utilized.

CHAPTER 6: CONCLUSION

We have developed AI models that can automatically predict treatment need based on IOTN-AC reference standards into two groups (binary) with up to 78% accuracy, 95% sensitivity, 82% specificity, 83% positive predictive value, and 88% negative predictive value, and into three groups (ternary) with up to 71% accuracy, 66% sensitivity, 82% specificity, 82% positive predictive value, and 84% negative predictive value. Results can presumably be improved with an increase in sample/training size. Accuracy decreases slightly with lack of overjet value. Future studies should be conducted with increased sample size. Potential uses for this AI system include in-office assessments by orthodontists, at-home assessments by patients, and utilization by third-party payers and the NHS.

Bibliography

1. Hung M, Su S, Hon ES, et al. Examination of orthodontic expenditures and trends in the United States from 1996 to 2016: disparities across demographics and insurance payers. *BMC Oral Health*. 2021;21(1):1-10. doi:10.1186/s12903-021-01629-6
2. Laniado N, Oliva S, Matthews GJ. Children's orthodontic utilization in the United States: Socioeconomic and surveillance considerations. *Am J Orthod Dentofac Orthop*. 2017;152(5):672-678. doi:10.1016/j.ajodo.2017.03.027
3. Shaw WC. The influence of children's dentofacial appearance on their social attractiveness as judged by peers and lay adults. *Am J Orthod*. 1981;79(4):399-415. doi:10.1016/0002-9416(81)90382-1
4. Papio MA, Fields HW, Beck FM, Firestone AR, Rosenstiel SF. The effect of dental and background facial attractiveness on facial attractiveness and perceived integrity and social and intellectual qualities. *Am J Orthod Dentofac Orthop*. 2019;156(4):464-474.e1. doi:10.1016/j.ajodo.2018.10.021
5. Olsen JA, Inglehart MR. Malocclusions and perceptions of attractiveness, intelligence, and personality, and behavioral intentions. *Am J Orthod Dentofac Orthop*. 2011;140(5):669-679. doi:10.1016/j.ajodo.2011.02.025
6. Trivedi K, Shyagali TR, Doshi J, Rajpara Y. Reliability of Aesthetic component of IOTN in the assessment of subjective orthodontic treatment need. *J Adv Oral Res*. 2011;2(1):59-66. doi:10.1177/2229411220110111
7. Richmond S, Shaw WC, O'Brien KD, et al. The relationship between the index of orthodontic treatment need and consensus opinion of a panel of 74 dentists. *Br Dent J*. 1995;178(10):370-374. doi:The relationship between the index of orthodontic treatment need and consensus opinion of a panel of 74 dentists
8. Younis J, Vig K, DJ R, RJ W. A validation study of three indexes of orthodontic treatment need in the United States. *Community Dent Oral Epidemiol*. 1997;25:358-362. doi:10.1111/j.1600-0528.1997.tb00955.x
9. Borzabadi-Farahani A. An Overview of Selected Orthodontic Treatment Need Indices. In: Naretto S, ed. *Principles in Contemporary Orthodontics*. InTech; 2011. doi:10.5772/19735
10. Stenvik A, Espeland L, Linge BO, Linge L. Lay attitudes to dental appearance and need for orthodontic treatment. *Eur J Orthod*. 1997;19:271-277. doi:0.1093/ejo/19.3.271
11. Evans R, Shaw WC. Preliminary evaluation of an illustrated scale for rating dental attractiveness. *Eur J Orthod*. 1987;9:314-318.
12. Richmond S, Shaw WC, O'Brien KD, et al. The relationship between the index of orthodontic treatment need and consensus opinion of a panel of 74 dentists. *Br Dent J*. 1995;178(10):370-374. doi:10.1038/SJ.BDJ.4808776
13. Üçüncü N, Ertugay E. The use of the index of orthodontic treatment need (IOTN) in a school population and referred population. *J Orthod*. 2001;28(1):45-52. doi:10.1093/ortho/28.1.45
14. Fields H, Van W, Vig K. Reliability of Soft Tissue Profile Analysis in Children. *Angle Orthod*. 1982;52:159-165.
15. Richmond S. *Evaluating Effective Orthodontic Care*. First Numerics Ltd; 2018.
16. Livas C, Delli K. Subjective and objective perception of orthodontic treatment need: A systematic review. *Eur J Orthod*. 2013;35(3):347-353. doi:10.1093/ejo/cjr142

17. Holmes A. The subjective need and demand for orthodontic treatment. *Br J Orthod*. 1992;19(4):287-297. doi:10.1179/bjo.19.4.287
18. *Needs Assessment for Orthodontic Services in London Needs Assessment for Orthodontic Services in London*.; 2015. www.gov.uk/phe%0Awww.facebook.com/PublicHealthEngland
19. Ko CC, Tanikawa C, Wu TH, et al. EMBRACING NOVEL TECHNOLOGIES IN DENTISTRY AND ORTHODONTICS. In: *Craniofacial Growth Series*. Vol 56. ; 2020:117-135.
20. Goodfellow I, Bengio Y, Couorville A. *Deep Learning*. MIT Press; 2017.
21. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digit Med*. 2021;4. doi:10.1038/s41746-021-00438-z
22. Shan T, Tay FR, Gu L. Application of Artificial Intelligence in Dentistry. *J Dent Res*. 2021;100(3):232-244. doi:10.1177/0022034520969115
23. Mohammad-Rahimi H, Nadimi M, Rohban MH, Shamsoddin E, Lee VY, Motamedian SR. Machine learning and orthodontics, current trends and the future opportunities: A scoping review. *Am J Orthod Dentofac Orthop*. 2021;160(2):170-192.e4. doi:10.1016/j.ajodo.2021.02.013
24. Mohammad-Rahimi H, Motamedian SR, Rohban MH, et al. Deep learning for caries detection: A systematic review. *J Dent*. 2022;122(January):104115. doi:10.1016/j.jdent.2022.104115
25. Miethke R. Possibilities and limitations of various cephalometric variables and analysis. In: *Orthodontic Cephalometry*. Mosby-Wolfee; 1995:63-103.
26. Lagravère MO, Low C, Flores-Mir C, et al. Intraexaminer and interexaminer reliabilities of landmark identification on digitized lateral cephalograms and formatted 3-dimensional cone-beam computerized tomography images. *Am J Orthod Dentofac Orthop*. 2010;137(5):598-604. doi:10.1016/j.ajodo.2008.07.018
27. Schwendicke F, Chaurasia A, Arsiwala L, et al. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. *Clin Oral Investig*. 2021;25(7):4299-4309. doi:10.1007/s00784-021-03990-w
28. Hwang HW, Park JH, Moon JH, et al. Automated identification of cephalometric landmarks: Part 2-Might it be better than human? *Angle Orthod*. 2020;90(1):69-76. doi:10.2319/022019-129.1
29. Muraev AA, Tsai P, Kibardin I, et al. Frontal cephalometric landmarking: humans vs artificial neural networks. *Int J Comput Dent*. 2020;23(2):139-148. http://www.ncbi.nlm.nih.gov/pubmed/32555767
30. Li P, Kong D, Tang T, et al. Orthodontic Treatment Planning based on Artificial Neural Networks. *Sci Rep*. 2019;9(1):1-9. doi:10.1038/s41598-018-38439-w
31. Murata S, Lee C, Tanikawa C, Date S. Towards a fully automated diagnostic system for orthodontic treatment in dentistry. *Proc - 13th IEEE Int Conf eScience, eScience 2017*. Published online 2017:1-8. doi:10.1109/eScience.2017.12
32. Murata S, Ishigaki K, Lee C, Tanikawa C, Date S, Yoshikawa T. Towards a Smart Dental Healthcare: An Automated Assessment of Orthodontic Treatment Need. *Healthinfo*. 2017;(c):35-39.
33. JONES N, MARKS R, RAMIREZ R, RÍOS-VARGAS M. 2020 Census Illuminates Racial and Ethnic Composition of the Country. U.S. Census Bureau. Published 2020.

Accessed December 18, 2020. https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html
34. Proffit WR, Fields HW, Moray LJ. Prevalence of malocclusion and orthodontic treatment need in the United States: estimates from the NHANES III survey. *Int J Adult Orthodon Orthognath Surg*. 1998;13(2):97-106.
35. Alhammadi MS, Halboub E, Fayed MS, Labib A, El-Saaidi C. Global distribution of malocclusion traits: A systematic review. *Dental Press J Orthod*. 2018;23(6):e1-e10. doi:10.1590/2177-6709.23.6.40.e1-10.onl
36. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. 2016;2016-Decem:770-778. doi:10.1109/CVPR.2016.90
37. Myrianthous G. Training vs Testing vs Validation Sets. Towards Data Science. Published 2021. Accessed December 15, 2022. https://towardsdatascience.com/training-vs-testing-vs-validation-sets-a44bed52a0e1
38. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems*. Vol 32. Curran Associates, Inc.; 2019. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
39. Moral P Del, Nowaczyk S, Pashami S. Why Is Multiclass Classification Hard? *IEEE Access*. 2022;10:80448-80462. doi:10.1109/ACCESS.2022.3192514