bioengineering

Article

Artificial Intelligence for Predicting the Aesthetic Component of the Index of Orthodontic Treatment Need

Leah Stetzel [1], Florence Foucher [1], Seung Jin Jang [1], Tai-Hsien Wu [1], Henry Fields [1], Fernanda Schumacher [2], Stephen Richmond [3] and Ching-Chang Ko [1,*]

[1] Division of Orthodontics, The Ohio State University, 305 W. 12th Avenue, Columbus, OH 43210, USA
[2] Division of Biostatistics, The Ohio State University, 1841 Neil Avenue, Columbus, OH 43210, USA
[3] Department of Orthodontics, Cardiff University, Heath Park, Cardiff CF14 4XY, UK
* Correspondence: ko.367@osu.edu; Tel.: +1-919-688-3146

Abstract: 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. The objective of this study is to use artificial intelligence (AI) to automate the AC assessment. A total of 1009 pre-treatment frontal intraoral photos with overjet values were collected. Each photo was graded by an experienced calibration clinician. The AI was trained using the intraoral images, overjet, and two other approaches. For Scheme 1, the training data were AC 1–10. For Scheme 2, the training data were either the two groups AC 1–5 and AC 6–10 or the three 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. The intra-rater reliability for the grader, using kappa, was 0.84 (95% CI 0.76–0.93). Scheme 1 had 77% sensitivity, 88% specificity, 82% accuracy, 89% PPV, and 75% NPV in predicting the binary groups. All other schemes offered poor tradeoffs. Findings after omitting overjet and dataset supplementation results were mixed, depending upon perspective. We have developed deep learning-based algorithms that can predict treatment need based on IOTN-AC reference standards; this provides an adjunct to clinical assessment of dental aesthetics.

Keywords: Index of Orthodontic Treatment Need; aesthetic component; artificial intelligence

Citation: Stetzel, L.; Foucher, F.; Jang, S.J.; Wu, T.-H.; Fields, H.; Schumacher, F.; Richmond, S.; Ko, C.-C. Artificial Intelligence for Predicting the Aesthetic Component of the Index of Orthodontic Treatment Need. *Bioengineering* **2024**, *11*, 861. https://doi.org/10.3390/bioengineering11090861

Academic Editors: Hang Nga Mai and Tim Joda

Received: 21 July 2024
Revised: 16 August 2024
Accepted: 20 August 2024
Published: 23 August 2024

1. 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 fiscal 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]. NHS spending on primary care orthodontic services is approximately GBP £250 million annually. (£: Great British Pound) [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 the scientific literature. In a study by Shaw [4], it was found that children's dental features affect a viewer's perceptions 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-based 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 IOTN and supported its use with international populations [7–9]. IOTN has two components: the dental health component (DHC) and the aesthetic component (AC).

The DHC consists of a five-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 "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 esthetic impairment and, consequently, the psychosocial need for orthodontic treatment [11].

With the use of a validation exercise, Richmond et al. [8] reported that IOTN-AC grades could be partitioned into three treatment-need subgroups: no need, borderline need, and definite treatment need. These reflect AC grades 1–4, 5–7, and 8–10, respectively, in this modified grouping [12].

IOTN is currently used by the National Health Service (NHS) to determine whether children qualify for covered orthodontic treatment. Patients with an 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 [13,14]. 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 within the NHS [15]. It is clear that the AC evaluation impacts the ability of patients to receive care.

With increasing demand for orthodontic care, reducing the workload of orthodontists and increasing at-home patient assessments are appealing ideas. The use of artificial intelligence (AI) has aided both the medical and the dental fields in diagnosis automation [16–19]. Orthodontics may be one of the dental specialties earliest in adapting AI into its practice [20] A systematic review and meta-analysis conducted in 2021 seeking to examine the accuracy of deep learning (a branch of AI which utilizes neural networks) in detecting landmarks on cephalometric radiographs demonstrated relatively high accuracy [21]. Others noted that deep learning performed similar to seasoned clinicians, and perhaps even better than inexperienced ones [22,23].

AI's ability to assess orthodontic treatment needs has been explored. In a study by Murata et al. [24], AI was able to classify patients into five 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 the current study, we propose 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 patients' frontal intraoral images with the corresponding IOTN-AC classification and overjet and develop a deep-learning based AI algorithm that could identify the IOTN-AC.

2. Materials and Methods

2.1. Data Collection

A total of 1009 intraoral images were gathered in a quota-sampling manner, such that they mirrored the U.S. population according to race and overjet values, the latter established from the epidemiological literature [25–27]. The 1009 intraoral images with a corresponding median overjet were assessed by an experienced calibration examiner. Each photo was assigned an AC score. After a two-week wash-out period, 200 randomly chosen images were tested for reliability using kappa. The 1009 images then served as our gold standard.

2.2. 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 at central incisors. The modules consist of a convolutional neural network (CNN) module, an overjet module, and an output module, corresponding to the two inputs (Figure 1).

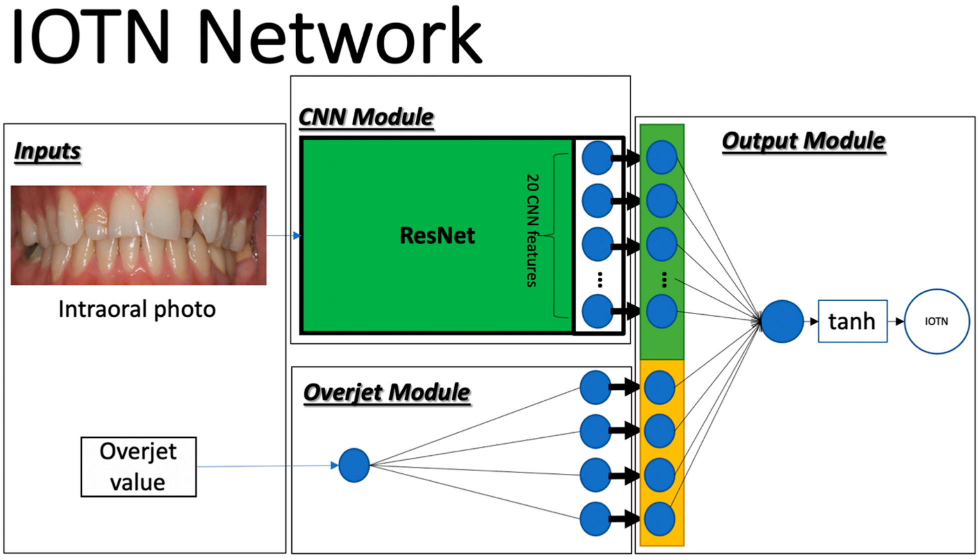

Figure 1. The IOTN network with 2 inputs, an overjet module, a CNN module, and an output module.

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 used to extract 20 hidden features [28]. 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 grade implies the severity of the patient's oral conditions. 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 the most used. 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) look more similar to those patients with IOTN 5–7 (borderline need for treatment) than to patients with IOTN 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 IOTN-AC grades are equidistant from each other, classifications can be considered interval data.

2.3. Implementation

After collection of the 1009 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 [29].

Of the 1009 gathered, 800 images were used in the training phase, 40 images were used in the validation phase, and 200 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 displaying how the IOTN network was trained.

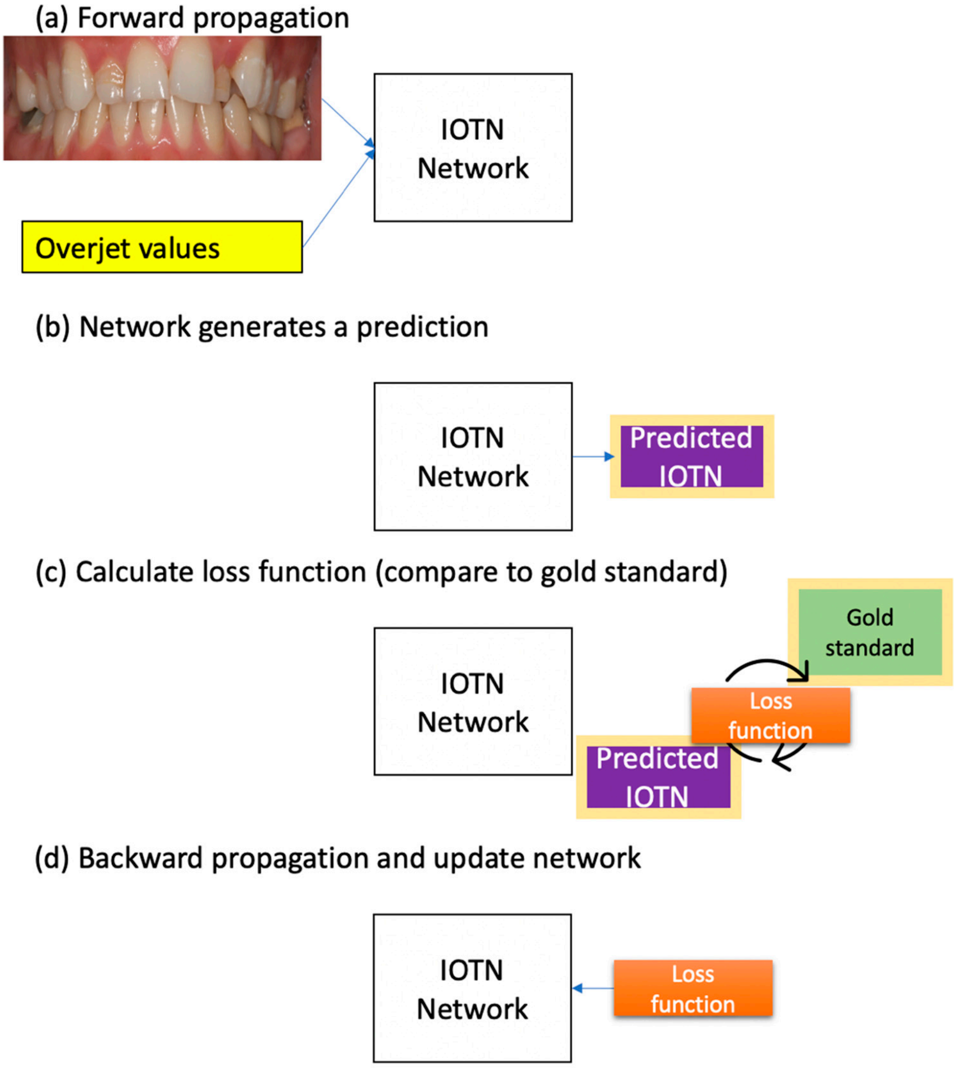

Figure 2. Schematic of a training process. (**a**) Forward propagation: The model receives the first two inputs, namely the intraoral image and the overjet value. (**b**) Network generates a prediction: The network generates a prediction based on these inputs, aiming to learn the output of the IOTN. (**c**) Calculate loss function: The predicted IOTN value is compared to the gold standard (third input) to calculate the discrepancy. (**d**) Backward propagation and update network: The discrepancy is back propagated through each layer of the network to update their parameters.

To test the model, the AI was given 200 unique new images with corresponding OJ values and was tasked with grading the IOTN-AC. Figure 3 shows a schematic of the testing phase. The testing dataset mirrored the IOTN distribution of our representative sample of 1009. That is, the testing data had the same percentage of each IOTN-AC grade as indicated in the initial data collection. Sensitivity (SEN) is the proportion of correct AI gradings for IOTN-AC images that were deemed as needing treatment by the gold standard. Specificity (SPE) is the proportion of correct AI gradings for images that were deemed as not needing treatment. Positive Predictive Value (PPV) is the proportion of AI gradings indicating a need for treatment that were correct. Negative Predictive Value (NPV) is the proportion of AI gradings indicating no need for treatment that were correct. Accuracy (ACC) is the overall proportion of correct AI gradings compared to the gold standard.

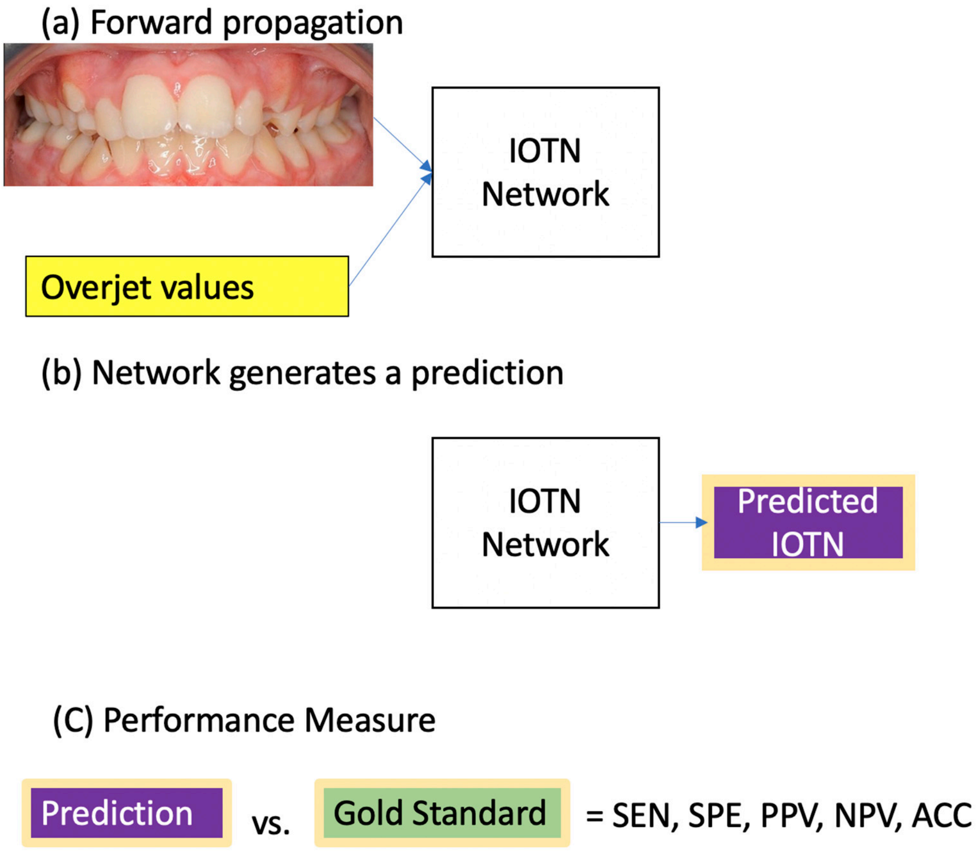

Figure 3. Schematic of a test process. (**a**) Forward propagation: The model takes the two inputs, namely the intraoral image and the overjet value. (**b**) Network generates a prediction: The network predicts an IOTN-AC grade based on these inputs. (**c**) Performance Measurement: Performance is assessed by comparing the predicted value to the gold standard using diagnostic metrics, including SEN (sensitivity), SPE (specificity), PPV (positive predictive value), NPV (negative predictive value), and ACC (accuracy).

2.4. 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 a random chance of being applied on the training images. By performing this heavy augmentation configuration, we expanded our training data to 200 images for each grade, for a total of 2000 images. It is important to note that although each grade was augmented in order to have 200 images, the image 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. Due to 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 as to its ability to predict IOTN. All the implementations were accomplished with Pytorch, an open-source deep learning library with the Python programming language [30]. 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.

1. Scheme 0

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

2. Scheme 1

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

3. 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.

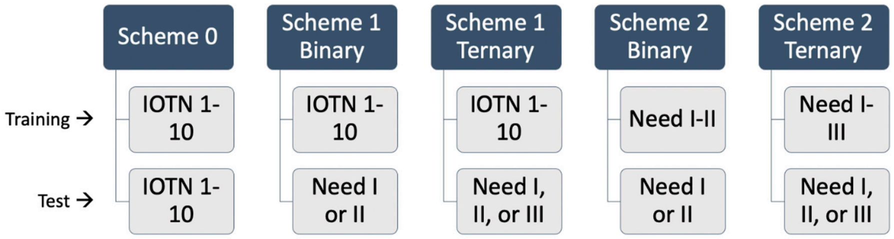

Figure 4. Summary of Scheme 0, Scheme 1, and Scheme 2: training and test. Need I or II is binary (IOTN 1–5 and 6–10). Need I, II, or III is ternary (IOTN 1–4, 5–7, and 8–10).

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 IOTN-AC were obtained from Dr. Richmond. The IOTN network was trained and tested again using Scheme 1.

2.5. Statistical Analysis

All schemes' performances were measured by calculating sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), accuracy (ACC), and balanced accuracy (BA) [31,32].

For the binary predictions, an IOTN-AC of 6–10 was considered a "positive" test and prediction, while an IOTN-AC of 1–5 was considered a "negative" test and prediction.

For the ternary prediction, sen, 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 IOTN 1–10, sen, spec, PPV, and NPV were calculated for each individual grade.

3. Results

The gold-standard IOTN grader demonstrated excellent intra-rater reliability in the identification of IOTN grades 1–10, as tested 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 U.S. population. The most infrequent IOTN-AC grade was IOTN 1, which represented 1% of our sample. IOTN-AC 9 and IOTN 10 were also uncommon, and each represented 3% of our sample. IOTN-AC 6 and 7 were the most frequent grades in our sample, representing 20% and 18%, respectively. A complete IOTN-AC distribution for our sample can be found in Table 1.

Table 1. IOTN-AC grades for our sample, which was selected in such a manner as to be representative of the IOTN-AC grades in the U.S. population.

IOTN	Count	Percentage
1	7	1%
2	49	5%
3	97	10%
4	134	13%
5	149	15%
6	203	20%
7	182	18%
8	125	12%
9	31	3%
10	32	3%

3.1. Prediction of IOTN-AC 1–10

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

3.2. Prediction of IOTN-AC 1–5 (I) and 6–10 (II)—Binary

For the binary predictions, Scheme 1 outperformed Scheme 2 in sensitivity, specificity, positive predictive value, negative predictive value, and accuracy. Scheme 1 was able to identify images with IOTN-AC 6–10 77% of the time. The results of the binary predictions for Scheme 1 and Scheme 2 can be seen in Figure 5.

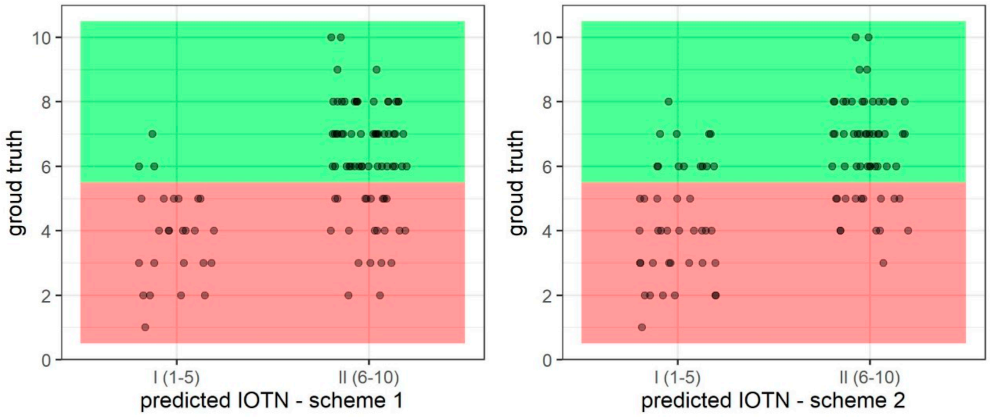

Figure 5. Scatter plots depicting binary prediction results for Scheme 1 (**left**) and Scheme 2 (**right**). In these plots, red region denotes IOTN-AC grades 1–5, while green region denotes IOTN-AC grades 6–10, as per the calibration clinician. Correct classifications are represented in the lower left and upper right quadrants of the plots.

3.3. Prediction of IOTN-AC 1–4 (I), 5–7 (II), and 8–10 (II)—Ternary

For the ternary predictions, on average, Scheme 1 outperformed Scheme 2 in all aspects: sensitivity, specificity, positive predictive value, negative predictive value, and accuracy.

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" categories, whereas Scheme 2 mis-predicted actual "Borderline Need" subjects by placing them 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 one of the actual "Great Need" subjects, placing them 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 are visualized in Figure 6.

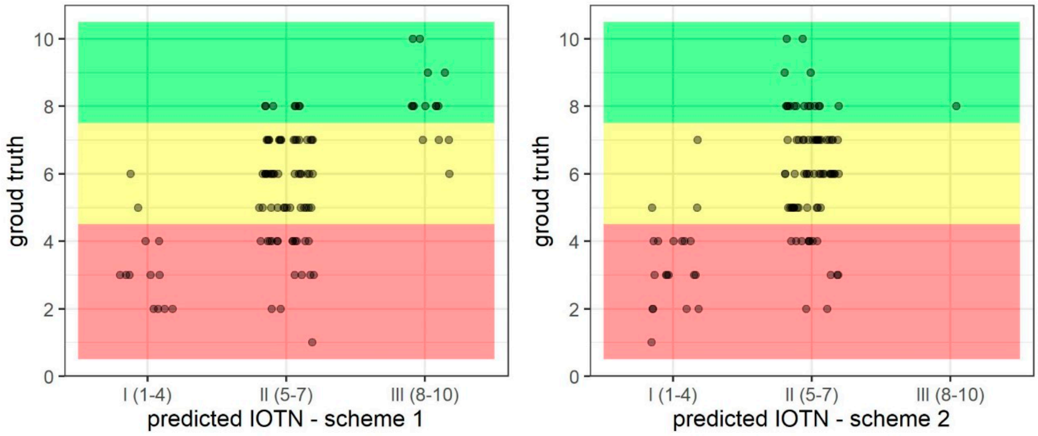

Figure 6. Scatter plots for ternary prediction results of Scheme 1 (**left**) and Scheme 2 (**right**). In these plots, red region denotes IOTN-AC grades 1–4, yellow region denotes IOTN-AC grades 5–7, and green region denotes IOTN-AC grades 8–10, as per the calibration clinician. Outcomes in the lower left, center, and upper right regions are correct classifications.

3.4. 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.

3.5. Predictions with Sample Size and Augmented Data

The model's accuracy was positively correlated with the increase in sample size; however, accuracy was substantially improved when up-sampling and image augmentation were implemented, increasing from 56% to 65% at a 25% sampling size. When increasing the sample size with image augmentation alone, the accuracy was saturated at 75%, but decreased by 15% at the end; this could be explained by the inability of the software to recognize the images once they were excessively tweaked. The results of the binary and ternary predictions with our supplemented data can be found in Figure 7.

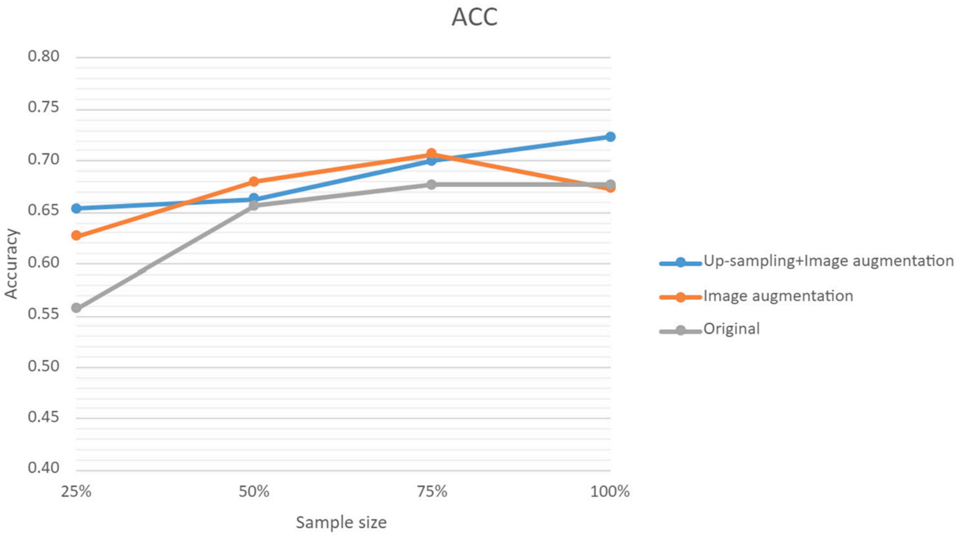

Figure 7. Line graph representing the accuracy of all schemes.

3.6. Summary

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

Table 2. Performance measures of the schemes.

Scheme	Sens	Spec	PPV	NPV	Acc
Scheme 0	0.27	0.92	0.50	0.92	0.34
Scheme 1 Binary	0.77	0.88	0.89	0.75	0.82
Scheme 2 Binary	0.76	0.87	0.88	0.74	0.81
Scheme 1 Ternary	0.65	0.83	0.77	0.85	0.72
Scheme 2 Ternary	0.63	0.81	0.67	0.82	0.67
Scheme 1 w/out OJ Binary	0.80	0.87	0.89	0.77	0.83
Scheme 1 w/out OJ Ternary	0.58	0.79	0.69	0.81	0.66

Sens, Sensitivity; Spec, specificity; PPV, positive predictive value; NPV, negative predictive value; Acc, accuracy. For Scheme 0 and any ternary predictions, sensitivity, specificity, positive predictive value, and negative predictive value were averaged.

4. Discussion

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

The experienced calibration examiner had nearly perfect intra-rater reliability by weighted kappa, according to the test described in Cohen [33]. 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 that this increased complexity is due, at least in part, to the heterogeneity in decision boundaries [34].

In order to improve our results, Scheme 1 and Scheme 2 were developed, and 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 promising among those 18 years or younger and enrolled in the NHS. If one is considered borderline in the DHC, the binary IOTN-AC classification can determine whether you are eligible for NHS-funded treatment (IOTN 6–10) or if you will be ineligible (IOTN 1–5). The ternary classification is more descriptive, in which IOTN 1–4 indicates little to no treatment need, IOTN 5–7 indicates moderate treatment need, and IOTN 8–10 indicates great treatment need, but is less useful in real application.

Certain metrics lead to the conclusion that Scheme 1 outperforms Scheme 2, and that Scheme 1 shows promise. 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's or a provider's viewpoint, it is undesirable to receive false negative scores. So, high sensitivity and high NPV are most important. You want all who deserve it to be funded, and do not care if there are some false positives, because all who qualify (and then some, possibly) will be funded. Our judgement is that with public funds, shortchanging those who qualify is worse than mistakenly funding a few extra cases in error. All deserving cases will be provided with funding.

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" [35]. Therefore, Scheme 1 with binary prediction could be considered promising for use by the NHS, after further improvement.

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 were to be 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 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 performs 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 captured.

Overall, the results of Scheme 1 were more promising than those of Scheme 2 when considering both binary and ternary predictions. Therefore, we decided to investigate how Scheme 1 would perform without overjet input. This would allow for less clinical error and less variation among practitioners. In the binary classification, ACC, SEN, and NPV increased by 1%, 3%, and 2%, respectively. There was a decrease in SPE, while PPV stayed the same. In the ternary classification, all values decreased slightly, while NPV improved. This slight decrease in results may not be clinically significant.

We investigated how increasing the sample or training sizes 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 U.S. population), IOTN grades 1, 9, and 10 were significantly under-represented. When the data were supplemented, the prevalence across classifications was increased, especially for IOTN 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.

Limitations of this study include the fact that while our AI system demonstrated quicker results compared to the manual method, it did not achieve acceptable accuracy. This limitation is considerable, as the typical threshold for acceptable accuracy in clinical dentistry and orthodontics is 90% or higher [36–38]. Our current model falls short of this benchmark, indicating the need for further refinement.

One significant challenge is that the current methodology may not fully leverage the potential of our AI system. There could be a discrepancy in how images are assessed; the AI may have failed to produce reliable results because it could not effectively capture the holistic approach that human observers use. Teeth can be viewed as a whole unit or a set of features. Understanding whether discrete changes in images correspond to a grading of severity or if severity is established more holistically by observers is crucial. Our current algorithm may not have effectively balanced these approaches, and future work should explore this aspect in greater depth.

Another limitation of our study is the relatively small sample size of approximately one thousand images, which may affect the robustness of our AI algorithm in diverse clinical scenarios. While our sample was designed to mirror the U.S. population according to race, the potential differences in algorithmic errors across different racial groups were not specifically investigated. The variability in gingival color among races might influence the performance of the AI model [39]. Investigating these differences is necessary to improve the algorithm's accuracy and fairness across diverse populations [40]. Additionally, the variability in image quality and the need for image manipulation may introduce errors. Standardized photography protocols could help reduce these errors by minimizing the need for extensive image augmentation. Implementing consistent imaging techniques might enhance the algorithm's accuracy and reliability. Furthermore, plans for future work include integrating this algorithm with image segmentation techniques to separate the teeth from the gingiva in intraoral photos, mitigating the influence of gingival color on the algorithm's predictions. Future studies with larger and more diverse datasets are necessary to thoroughly evaluate these aspects and to determine the impacts of racial differences and standardized photography on algorithm accuracy.

In summary, while our study provides insights into developing an AI solution for the aesthetic component of the IOTN, further work remains in order to achieve the required accuracy for clinical usability. Future studies should focus on refining the algorithm, balancing the training dataset, and exploring both holistic and analytic image assessment approaches to enhance the system's reliability and applicability in a clinical setting.

5. Conclusions

We have developed deep learning-based algorithms capable of predicting dental aesthetic needs based on IOTN-AC reference standards. Our approach, using the AC 1–10 scale input with binary testing, proved superior compared to other AC categorizations, aligning well with patient-centered public policy perspectives. While our AI system has shown promising results, achieving a level of accuracy that is not yet sufficient for clinical use highlights the potential for further refinement. Specifically, the elimination of overjet enhanced and simplified our results, demonstrating potential improvements in accuracy and clinical applicability. Further studies should focus on refining the algorithm and exploring holistic versus analytic approaches to image assessment in order to enhance reliability in clinical settings.

Author Contributions: Conceptualization, H.F., C.-C.K. and L.S.; methodology, T.-H.W., F.S. and L.S.; software, T.-H.W.; validation, L.S., F.F. and S.J.J.; formal analysis, T.-H.W., F.S.; investigation, C.-C.K., S.R. and T.-H.W.; resources, C.-C.K.; data curation, L.S., F.F. and S.R.; writing—review and editing, L.S., H.F., F.F., S.J.J. and C.-C.K.; supervision, C.-C.K. and H.F.; project administration, C.-C.K. and T.-H.W.; funding acquisition, C.-C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ching-Chang Ko's start-up at OSU (G100125).

Institutional Review Board Statement: The study was reviewed by the OSU IRB committee, which determined that the project was exempt from IRB review.

Informed Consent Statement: Patient consent was waived due to the nature of the retrospective study by chart review as well as the deidentified data collection.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. An NHS under Pressure. BMA, 2022. Available online: https://www.bma.org.uk/advice-and-support/nhs-delivery-and-workforce/pressures/an-nhs-under-pressure (accessed on 15 February 2023).
2. The 'Zoom Boom': BOS Stats Reveal a Surge in Demand for Orthodontics during the Pandemic. British Orthodontic Society, 2021. Available online: https://www.dentalreview.news/dentistry/21-orthodontic/7269-bos-reports-zoom-boom-surge-in-orthodontics (accessed on 15 February 2023).
3. Price, J.; Whittaker, W.; Birch, S.; Brocklehurst, P.; Tickle, M. Socioeconomic disparities in orthodontic treatment outcomes and expenditure on orthodontics in England's state-funded National Health Service: A retrospective observational study. *BMC Oral Health.* **2017**, *17*, 123. [CrossRef] [PubMed]
4. Shaw, W.C. The influence of children's dentofacial appearance on their social attractiveness as judged by peers and lay adults. *Am. J. Orthod.* **1981**, *79*, 399–415. [CrossRef] [PubMed]
5. Papio, M.A.; Fields, H.W.; Beck, F.M.; Firestone, A.R.; Rosenstiel, S.F. 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*, 464–474.e1. [CrossRef]
6. Olsen, J.A.; Inglehart, M.R. Malocclusions and perceptions of attractiveness, intelligence, and personality, and behavioral intentions. *Am. J. Orthod. Dentofac. Orthop.* **2011**, *140*, 669–679. [CrossRef]
7. Trivedi, K.; Shyagali, T.R.; Doshi, J.; Rajpara, Y. Reliability of Aesthetic component of IOTN in the assessment of subjective orthodontic treatment need. *J. Adv. Oral. Res.* **2011**, *2*, 59–66. [CrossRef]
8. Richmond, S.; Shaw, W.C.; O'Brien, K.D.; Buchanan, I.B.; Stephens, C.D.; Andrews, M.; Roberts, C.T. The relationship between the index of orthodontic treatment need and consensus opinion of a panel of 74 dentists. *Br. Dent. J.* **1995**, *178*, 370–374. [CrossRef] [PubMed]
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*, 358–362. [CrossRef] [PubMed]
10. Borzabadi-Farahani, A. An Overview of Selected Orthodontic Treatment Need Indices. In *Principles in Contemporary Orthodontics*; Naretto, S., Ed.; InTech: Rijeka, Croatia, 2011; Available online: http://www.intechopen.com/books/principles-in-contemporary-orthodontics/an-overview-of-selected-orthodontic-treatment-need-indices (accessed on 21 August 2024).
11. Stenvik, A.; Espeland, L.; Linge, B.O.; Linge, L. Lay attitudes to dental appearance and need for orthodontic treatment. *Eur. J. Orthod.* **1997**, *19*, 271–277. [CrossRef] [PubMed]
12. 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*, 45–52. [CrossRef]
13. Livas, C.; Delli, K. Subjective and objective perception of orthodontic treatment need: A systematic review. *Eur. J. Orthod.* **2013**, *35*, 347–353. [CrossRef]
14. Holmes, A. The subjective need and demand for orthodontic treatment. *Br. J. Orthod.* **1992**, *19*, 287–297. [CrossRef] [PubMed]
15. *Needs Assessment for Orthodontic Services in London Needs Assessment for Orthodontic Services in London*; Public Health England: London, UK, 2015. Available online: https://www.gov.uk/government/publications/orthodontic-services-in-london-needs-assessment (accessed on 21 August 2024).
16. Shan, T.; Tay, F.R.; Gu, L. Application of Artificial Intelligence in Dentistry. *J. Dent. Res.* **2021**, *100*, 232–244. [CrossRef]
17. Mohammad-Rahimi, H.; Nadimi, M.; Rohban, M.H.; Shamsoddin, E.; Lee, V.Y.; Motamedian, S.R. Machine learning and orthodontics, current trends and the future opportunities: A scoping review. *Am. J. Orthod. Dentofac. Orthop.* **2021**, *160*, 170–192.e4. [CrossRef] [PubMed]
18. Mohammad-Rahimi, H.; Motamedian, S.R.; Rohbanm, M.H.; Krois, J.; Uribe, S.E.; Mahmoudinia, E.; Rokhshad, R.; Nadimi, M.; Schwendicke, F. Deep learning for caries detection: A systematic review. *J. Dent.* **2022**, *122*, 104115. [CrossRef]
19. Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D.S.W.; Karthikesalingam, A.; King, D.; Ashrafian, H.; Darzi, A. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *NPJ Digit. Med.* **2021**, *4*, 65. [CrossRef]
20. Ko, C.-C.; Tanikawa, C.; Wu, T.-H.; Pastewait, M.; Bonebreak Jackson, C.; Kwon, J.J.; Lee, Y.-T.; Lian, C.; Wang, L.; Shen, D. Embracing Novel Technologies in Dentistry and Orthodontics. In *Craniofacial Growth Series*; University of Michigan: Ann Arbor, MI, USA, 2020; pp. 117–135.
21. Schwendicke, F.; Chaurasia, A.; Arsiwala, L.; Lee, J.H.; Elhennawy, K.; Jost-Brinkmann, P.G.; Demarco, F.; Krois, J. Deep learning for cephalometric landmark detection: Systematic review and meta-analysis. *Clin. Oral Investig.* **2021**, *25*, 4299–4309. [CrossRef]
22. Hwang, H.W.; Park, J.H.; Moon, J.H.; Yu, Y.; Kim, H.; Her, S.B.; Srinivasan, G.; Aljanabi, M.N.A.; Donatelli, R.E.; Lee, S.-J. Automated identification of cephalometric landmarks: Part 2-Might it be better than human? *Angle Orthod.* **2020**, *90*, 69–76. [CrossRef]
23. Muraev, A.A.; Tsai, P.; Kibardin, I.; Oborotistov, N.; Shirayeva, T.; Ivanov, S.; Guseynov, N.; Aleshina, O.; Bosykh, Y.; Safyanova, E.; et al. Frontal cephalometric landmarking: Humans vs artificial neural networks. *Int. J. Comput. Dent.* **2020**, *23*, 139–148. [PubMed]
24. 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**, *2017*, 35–39.

25. Jones, N.; Marks, R.; Ramirez, R.; Ríos-Vargas, M. 2020 Census Illuminates Racial Ethnic Composition of the Country, U.S. Census Bureau. 2020. Available online: https://biz.libretexts.org/Courses/Folsom_Lake_College/BUS_330:_Managing_Diversity_in_the_Workplace_(Buch)/Unit_3:_Diversity_Groups_and_Categories:_Part_B/Chapter_6:_European_Americans_Native_Americans_and_Multi-racial_Americans/6.3:_Multi-racial_Americans/6.3.1:_2020_Census_Illuminates_Racial_and_Ethnic_Composition_of_the_Country (accessed on 18 December 2020).
26. Proffit, W.R.; Fields, H.W.; Moray, L.J. 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*, 97–106. [PubMed]
27. Alhammadi, M.S.; Halboub, E.; Fayed, M.S.; Labib, A.; El-Saaidi, C. Global distribution of malocclusion traits: A systematic review. *Dental. Press J. Orthod.* **2018**, *23*, 40.e1. [CrossRef]
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Myrianthous, G. Training vs. Testing vs. Validation Sets. Towards Data Science, 2021. Available online: https://towardsdatascience.com/training-vs-testing-vs-validation-sets-a44bed52a0e1 (accessed on 15 December 2022).
30. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2019; Available online: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf (accessed on 21 August 2024).
31. Trevethqn, R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Front. Public Health* **2017**, *5*, 307.
32. Kelleher, J.; Namee, B.; D'Arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics*; The MIT Press: Cambridge, MA, USA, 2015.
33. McHugh, M.L. Lessons in biostatistics interrater reliability: The kappa statistic. *Biochem. Medica.* **2012**, *22*, 276–282. [CrossRef]
34. Del Moral, P.; Nowaczyk, S.; Pashami, S. Why Is Multiclass Classification Hard? *IEEE Access* **2022**, *10*, 80448–80462. [CrossRef]
35. The NHS Constitution. Available online: https://www.gov.uk/government/publications/the-nhs-constitution-for-england/the-nhs-constitution-for-england (accessed on 18 February 2023).
36. Lee, J.-H.; Kim, D.H.; Jeong, S.N.; Choi, S.H. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. *J. Dent.* **2018**, *77*, 106–111. [CrossRef]
37. Chen, H.; Zhang, K.; Lyu, P.; Li, H.; Zhang, L.; Wu, J.; Lee, C.H. A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films. *Sci. Rep.* **2019**, *9*, 3840. [CrossRef]
38. Yu, H.J.; Cho, S.; Kim, M.; Kim, W.; Kim, J.; Choi, J. Automated Skeletal Classification with Lateral Cephalometry Based on Artificial Intelligence. *J. Dent. Res.* **2020**, *99*, 249–256. [CrossRef]
39. Koppolu, P.; Almutairi, H.; al Yousef, S.; Ansary, N.; Noushad, M.; Vishal, M.B.; Swapna, L.A.; Alsuwayyigh, N.; Albalawi, M.; Shrivastava, D.; et al. Relationship of skin complexion with gingival tissue color and hyperpigmentation. A multi-ethnic comparative study. *BMC Oral Health* **2024**, *24*, 451. [CrossRef]
40. Ramirez, A.H.; Sulieman, L.; Schlueter, D.J.; Halvorson, A.; Qian, J.; Ratsimbazafy, F.; Loperena, H.R.; Mayo, K.; Basford, M.; Deflaux, N.; et al. The *All of Us* Research Program: Data quality, utility, and diversity. *Patterns* **2022**, *3*, 100570. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.