

COMPARATIVE ANALYSIS OF THREE MACHINE LEARNING MODELS FOR EARLY PREDICTION OF SKELETAL CLASS-III MALOCCLUSION FROM PROFILE PHOTOS

Selahattin AKSOY, Department of Bioinformatics, Muğla Sıtkı Koçman University, Turkey, selahattinaksoy@mu.edu.tr
([ORCID](https://orcid.org/0000-0002-2574-7237))

Banu KILIÇ, Department of Orthodontics, Bezmialem Vakıf University, Turkey, bkilic@bezmialem.edu.tr
([ORCID](https://orcid.org/0000-0002-9207-4490))

Tuğba ÖNAL SÜZEK*, Department of Bioinformatics, Muğla Sıtkı Koçman University, Turkey, tugbasuzek@mu.edu.tr
([ORCID](https://orcid.org/0000-0002-3243-1759))

Received: 25.04.2022, Accepted: 14.08.2022

*Corresponding author

Research Article

DOI: 10.22531/muglajsci.1108397

Abstract

The pre-adolescent growth period is the best time for the skeletal Class-III malocclusion treatment. Diagnosis and treatment during this period continue to be a complex orthodontic problem. Class-III malocclusion is complicated to treat with braces frequently requiring surgical intervention after a pubertal growth spurt. In addition, delayed recognition of the problem will yield significant functional, aesthetic, and psychological concerns. This study presents the first fully automated machine learning method to accurately diagnose Class-III malocclusion applied across mobile images, to the best of our knowledge. For this purpose, we comparatively evaluated three machine learning approaches: a deep learning algorithm, a machine learning algorithm, and a rule-based algorithm. We collected a novel profile image data set for this analysis along with their formal diagnosis from 435 orthodontics patients. The most successful method among the three was the machine learning method, with an accuracy of %76.

Keywords: Class-III Malocclusion, machine learning, medical informatics.

PROFİL FOTOĞRAFLARINDAN İSKELETSEL SINIF III MALOKLÜZYONUN ERKEN TAHMİNİ İÇİN ÜÇ MAKİNE ÖĞRENİMİ MODELİNİN KARŞILAŞTIRMALI ANALİZİ

Özet

İskeletsel deformiteleri içinde Sınıf III maloklüzyonların tedavisi için en uygun zaman preadölesan dönemdir. Bu dönemde teşhis ve tedavi karmaşık bir ortodontik problem olmaya devam etmektedir. Sınıf III maloklüzyonun, pubertal büyüme atağı sonrası sıklıkla cerrahi müdahale gerektiren diş telleri ile tedavi edilmesi özellikle zordur. Ayrıca problemin geç fark edilmesi önemli fonksiyonel, estetik ve psikolojik kaygılara yol açacaktır. Çalışmamızda derin öğrenme algoritması, makine öğrenme algoritması ve kural tabanlı bir algoritma olmak üzere üç öğrenme modelinin doğruluklarını karşılaştırmalı analiz ettik. Bu analiz için, 435 ortodonti hastasından resmi teşhisleri ile birlikte özgün bir profil görüntüsü veri seti topladık. Bu üç yöntem arasında en başarılı olanı %76 doğruluk oranı ile makine öğrenmesi yöntemi olmuştur.

Anahtar Kelimeler: Sınıf III maloklüzyon, makine öğrenmesi, tıbbi bilişim.

Cite

Aksoy, S., Kılıç, B., Süzek, T. Ö., (2022). "Comparative Analysis of Three Machine Learning Models for Early Prediction of Skeletal Class-III Malocclusion from Profile Photos", *Mugla Journal of Science and Technology*, 8(2), 22-30.

1. Introduction

Skeleton Class-III malocclusion is characterized by the sagittal developmental retardation and/or rearward positioning of the maxilla or the extreme development and/or forward positioning of the mandible [1]. Class-III malocclusion causing a concave profile causes aesthetic anxiety in the child and the parents [2]. The correction of the harmony between the mandible and maxilla can be

provided by orthodontic/orthopedic appliances to be performed in early terms on the patients affecting the active sutures unclosed. Although the ideal age of treatment with orthopedic devices is preferably eight at the latest, it has also been reported that they can be effective up to 12 years of age [3]. It has been shown that early Class-III orthopedic treatment reduces the need for orthognathic surgery [4]. The American Association of Orthodontists (AAO) recommends that all children be

checked by an orthodontist no later than age seven. However, the most common age group during which patients seek orthodontic treatment is 12 years and older [5]. Moreover, many children do not have access to an orthodontic pre-examination at seven and before. Advances in consumer electronics and portable communications systems, particularly mobile phones, have led to faster and less expensive approaches to developing Point of Care Diagnosis [6]. Number of mobile subscribers globally according to GSMA Intelligence 2021 data reached approximately 5.22 billion [7]. Triggered by the Covid-19 pandemic period, the telehealth industry and more specifically the mobile health industry has become a 100-billion-dollar market by increasing five times since 2016 as of the end of 2021 [8]. There are several dental mobile applications in the current healthcare mobile application market [9]. Current applications are mainly developed for patient education about general dentistry [10], braces, Invisalign (Align Technology, San Jose, Calif), and oral health [11]. The commercial software Phimentum [12] claims using deep-learning for automatic landmark point detection on cephalometric images. Unlike Phimentum, our program has the advantage of not requiring cephalometric image input and/or an orthodontist's supervised selection of landmark points can compute the probability of Class-III diagnosis without any other apparatus other than mobile phones.

Our previous study [13], implemented an unsupervised diagnostic method based on the angles computed for Turkish adult patients [14]. Our previous study's disadvantage was that it was only applicable to Turkish adult patients. Our present study aimed to compare the accuracies of three alternative unsupervised machine learning approaches for malocclusion diagnosis. Best performing model of three will be adaptable to be trained on any given image set of any race or age rather than a fixed threshold determined for Turkish patients. For this purpose, we implemented and presented the accuracies of three methods; deep-learning algorithm, a machine learning algorithm, and a rule-based algorithm for an in-house image data set we collected from orthodontics patients.

2. Collection of Data

The data set created within the project's scope consists of profile photos of patients visiting Orthodontics at Bezmialem Vakıf University Faculty of Dentistry. The Institutional Review Board approval (IRB# 54022451-050.05.04) was obtained from Bezmialem Vakıf University to use profile image data of patients who applied to Bezmialem Vakıf University in our project. The orthodontist with 20 years of orthodontics clinical experience diagnosed the patient images and classified them into three classes, Class-I, II, and III using Dolphin

Imaging software (Version 11.95). The profile image taken for each patient is saved with an anonymized name in JPG/PNG format. Our anonymized dataset consists of 435 profile images and with their formal diagnosis from the orthodontics clinic.

3. Materials and Methods

Three alternative methods, namely the Rule-based, Deep Learning, and Machine Learning methods were implemented (Fig-1).

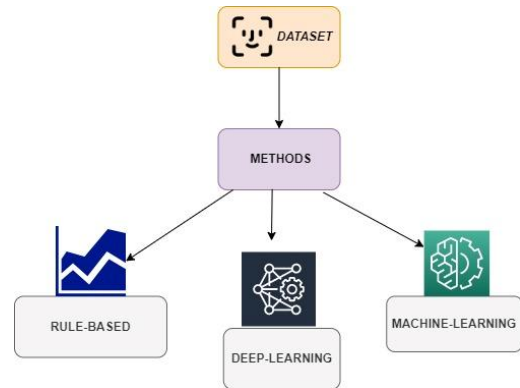


Figure 1. Flow-chart of the implemented pipeline

In these three methods we used mainly python programming language and its libraries scikit-learn for Machine Learning approach, OpenCV for image processing and pandas, seaborn etc. libraries for visualization and data analysis. In addition to that we used Tensorflow-Keras for deep learning approach. We used script base compiler to run our python codes and other required python libraries.

3.1. Deep Learning Approach

For Deep Learning Method, firstly, the images were pre-processed, and then the deep learning model was created.

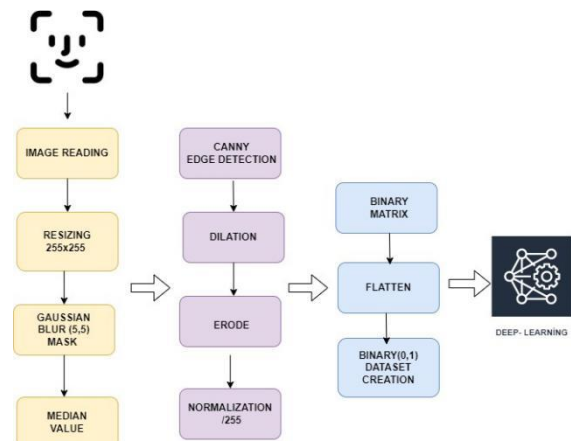


Figure 2. Image processing pipeline.

The images were first converted to gray-scale using the Python Open-CV library. Each image was scaled to 255X255 to standardize the images. Gaussian filtering was applied to scaled images using a 5x5 kernel matrix using OpenCV. Gaussian filtering (Gaussian Blur) was

applied to reduce the noise and detail. Then, the median values were computed for each picture. With these median values, the contours of the profile images were extracted using OpenCV "Canny Edge detection" on each picture. Using dilation (Spreading and Expansion) and erosion operations pixels were added to the silhouette to make them more prominent. We applied dilation to expand the borders; thanks to this expansion, we enlarged the pixel groups and reduced the spaces between the pixels.

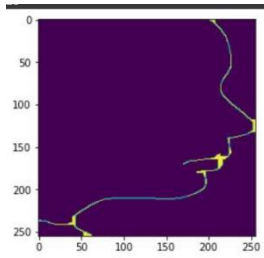


Figure 3. Silhouette output of the deep-learning pre-processing.

Next, the images were normalized using the min-max normalization formula (Fig-4), as all the images have a matrix with pixel values (0,255). As a result of this process, we converted our pictures into binary (0,1) form. Afterwards, we converted our matrix into a one-dimensional list by applying Flatten in order to use the 2D picture matrices we obtained in deep learning. As a result, each picture's silhouette, (Fig-3) was acquired in a single dimension.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Figure 4. Normalization formula.

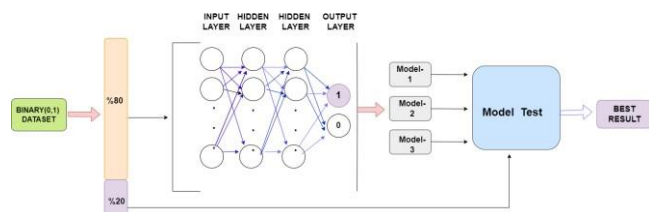


Figure 5. Deep learning model used.

The images were split into 80% training and 20% test set. 80% of the images were used for training the deep-learning model. Our deep learning model used a 4- layer structure, an input layer, two hidden layers, and an output layer. Since we were aiming a binary classification, we used "Relu" as the activation function in our input layer, and we used the sigmoid function in the intermediate layers. We used "Adam" as the optimizer and "cross entropy" as the optimizer loss function.

After creating the training model, we tested our model with our independent test data. As a result, a deep-learning model with 70% accuracy and 64% sensitivity was created for the disease classification. As a higher accuracy was required for a clinical setting, we experimented with additional unsupervised methods

that do not require a pre-set angle coefficient for diagnosis. To achieve this goal, we implemented the rule-based method as the next step.

Table 1. Accuracy of the deep learning method.

Accuracy	0.709
Precision	0.647
F1	0.578
Recall	0.523
AUC	0.673

3.2. Rule Based Approach

The same re-sizing and pre-processing steps were carried out for the critical value method on the images. Using the Python face-alignment library [15], 68 facial landmark points were selected., Po', Sn', A', Ls, Li, B, Pg', and Gn' were determined among these landmarks. After determining the other points requested by the orthodontist (Fig-6), the area between the Po' point at ear level and the area surrounding the Li (Lower Lip) part green area in Fig-7) and the remaining area between Ls - Sn'- Po' (red zone in Fig-7) Python OpenCV After masking, the ratio of the regions was computed in pixels.

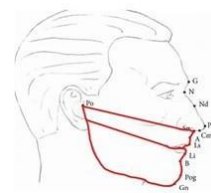


Figure 6. Two regions compared.



Figure 7. Two regions compared.

Table 2. Descriptions of facial landmarks.

Abbreviations	Definitions
N'	Soft tissue Nasion. Midpoint on the soft tissue contour of the base of the nasal root at the level of the frontonasal suture.
Ls	Labialis superior; the most anterior point of the upper lip.
Li	Labialis inferius: The most anterior point of the lower lip
Pg'	Soft tissue pogonion. The most anterior point on the mandible in the midline; the most anterior, prominent point on the chin.
Po'	Soft tissue porion : most superior point of the external acoustic meatus.
Pn	Pronasale; The tip of the nose.
Sn'	Soft tissue subnasal: The point at which thenasal septum merges with the uppercutaneous lip in the midsagittal plane.

A'	Soft tissue A point; The outer point of intersection between the A point horizontal line and the soft tissue.
B'	Soft tissue B point; The outer point of intersection between the B point horizontal line and the soft profile.
Soft Tissue Facial Plane	Line between soft tissue nasion to soft tissue pogonion
H line	Harmony line: drawn tangent to the soft tissue chin and the upper lip.
H angle	The angle formed between the softtissue facial plane line and the H line

For the area ratios of all patients, a total of 435 images, 162 Class-I, 163 Class-II, and 110 Class-III, we first plotted the descriptive statistical analysis and visualization.

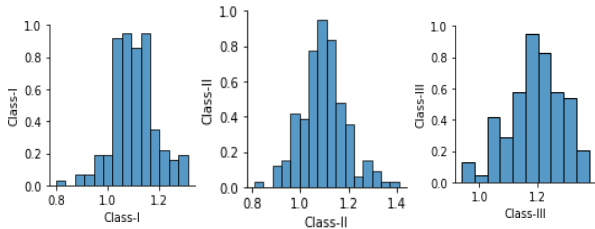


Figure 8. Histogram of each class showing the normalized number of samples.

Histogram graphs were plotted to show the distribution of each class; Class-I, Class-II, and Class-III (Fig-8). As seen in the histograms there is no left or right skewness (Left-Right Skewed Distribution) and the data show a normal distribution

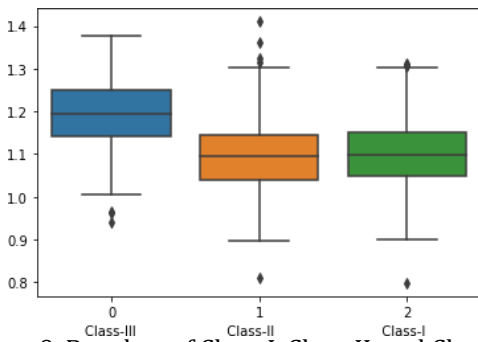
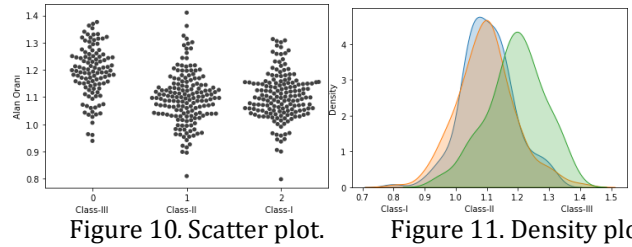


Figure 9. Boxplots of Class-I, Class-II, and Class-III.

The boxplots of the three classes in Fig-9 imply that the area ratios of Class-III patients were distinctly different from the other classes. Outliers in the boxplot were filtered and checked for misdiagnosis.

In Fig-10, scatter plot of the patients also shows that Class-III patients have distinct area ratio values. As seen in the density plot in Fig-11, the area values showed a normal distribution for all classes yet the mean of the Class-III area ratio was higher than the other two classes.



3.2.1. Evaluation of Class-III vs not

Classification of area ratios for our goal, we designed binary classification experiment of Class-III and not Class-III. In this setting, 110 patients were Class-III and 325 patients were not Class-III (Fig-12).

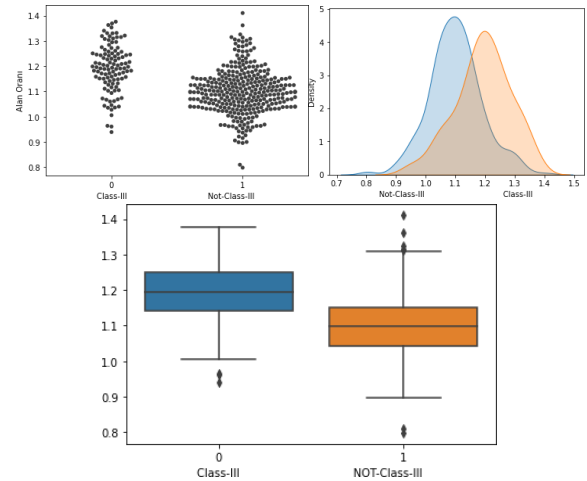


Figure 12. Distributions of Class-III and non-Class-III.

The area ratios of Class-III and non-Class-III followed a normal distribution (Figure 12-13).

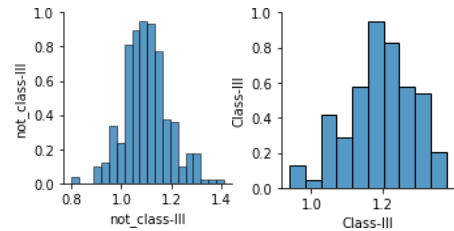
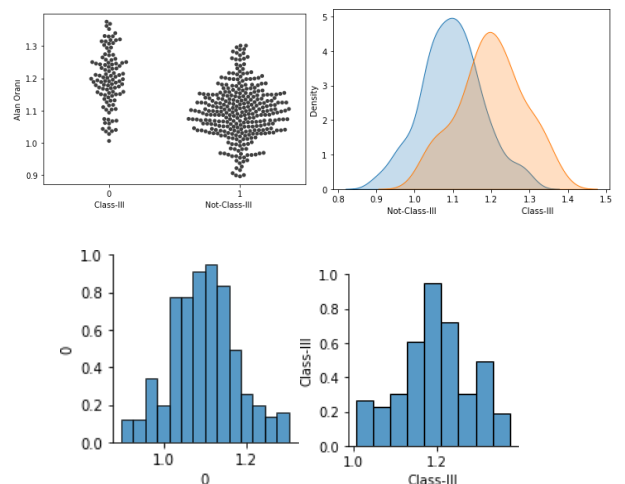


Figure 13. Histograms of Class-III and non-Class-III.

3.2.2. Cleaning Up Outliers



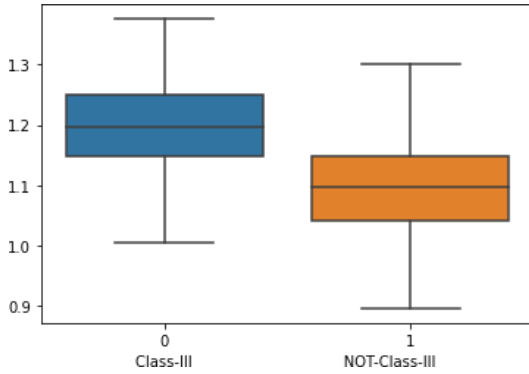


Figure 14. Plots of Class-III and non-Class-III after cleaning outliers.

Outliers were detected using the formula of Interquartile range (IQR) $[Q1 - 1.5 * IQR - Q3 + 1.5 * IQR]$, we found 12 outliers and removed them from the dataset. We obtained a total of 423 patients. We repeated the analysis as in Fig-14.

3.2.3. Experimental Setup

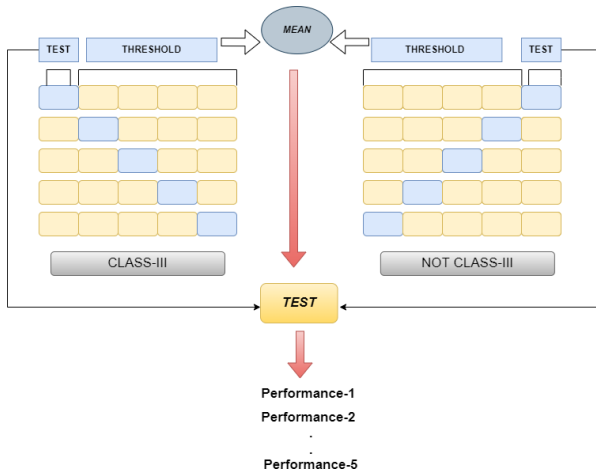


Figure 15. Experimental set up for classification.

To check whether the distinct area ratio difference of Class-III is random or not, we have created an experimental setup (Fig-15) in the form of a critical value (threshold). In the experiment setup, we divided the Class-III and non-Class-III data into 5 parts with 5-fold separately, and the average of 4 parts of the data set was labeled training. The same process was repeated for the non-Class-III dataset. The resulting rule-based value was tested with the independent test dataset of Class-III and non-Class-III data. The process was repeated for the entire 5-fold. The classification was labeled not Class-III if the test data area ratio was greater than the critical value and Class-III if it was less than the critical value. As a result of classification

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$F1 = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

$$\text{Specificity} = TN / (TN + FP)$$

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN)$$

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

3.2.4. Results

Table 3. The result of the experiment (3.2.3) for the whole data set.

	Accuracy	Recall	Precision	F1	Specificity
Fold- 1	0.712	0.636	0.451	0.528	0.738
Fold- 2	0.735	0.727	0.484	0.581	0.738
Fold- 3	0.666	0.772	0.414	0.539	0.630
Fold- 4	0.758	0.863	0.513	0.644	0.723
Fold- 5	0.793	0.727	0.571	0.640	0.815
Mean	0.733	0.745	0.487	0.586	0.729
STD	0.042	0.073	0.053	0.048	0.058

We shuffled our experimental dataset 100 times to avoid bias, dividing the dataset 5-fold each time. The average and standard deviation of these 100 training shuffles were computed below.

Table 4. Summary of the mean accuracies of 100 repeats.

	Accuracy	Recall	Precision	F1	Specificity
Mean	0.734	0.749	0.486	0.587	0.729
STD	0.035	0.092	0.046	0.050	0.050

3.2.5. Results (without outliers)

The same calculation and method were also applied on the data without the outliers, which increased the performance by ~%2.

After clearing the outliers we randomly shuffled our dataset each time, divided it into 5-Folds, repeated it 100 times, and calculated the average and standard deviation of these 100 training tests so that there would be no bias during training and testing.

Table 5. The result of the experiment (3.2.3) for the cleaned dataset.

	Accuracy	Recall	Precision	F1	Specificity
Fold- 1	0.755	0.818	0.514	0.631	0.734
Fold- 2	0.729	0.727	0.484	0.581	0.730
Fold- 3	0.750	0.571	0.500	0.533	0.809
Fold- 4	0.690	0.714	0.428	0.535	0.682
Fold- 5	0.809	0.809	0.586	0.680	0.809
Mean	0.747	0.728	0.502	0.592	0.753
STD	0.038	0.088	0.0508	0.056	0.049

Table 6. Summary of the mean accuracies of 100 repeats.

	Accuracy	Recall	Precision	F1	Specificity
Mean	0.748	0.750	0.503	0.599	0.747
STD	0.047	0.094	0.066	0.064	0.059

According to the results of the area ratio rule based experiment, specific facial landmarks such as Po', Sn', A', Ls, Li, B', Pog', Gn', the area surrounding the Po' point at the level of the Li (Lower Lip) and the remaining area between Li' - Sn' - Po' (red zone in Fig-7) differs distinctly for skeletal Class-III malocclusion patients. The ratio method is an important distinguishing factor for Class-III classification; therefore, we decided to add it as a feature to our machine learning model.

3.3. Machine Learning Approach

The machine learning method was performed in two main steps.

- 3.3.1) Extraction of orthodontic features from images.
- 3.3.2) The data set was labeled and the machine learning method is applied.

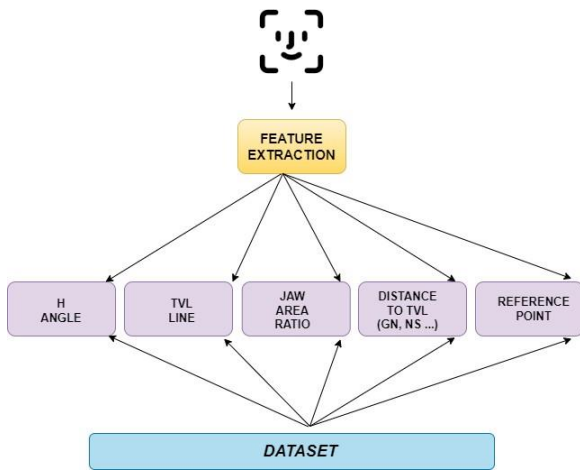


Figure 16. Extraction of orthodontic features from images.

3.3.1. Feature Extraction from Images

For calculation of H-Angle, we used the Python face-alignment library [15]. Firstly soft tissue nasion(N'), soft tissue chin(Pg') and upper lip(Ls) points (see Table 2) were computed. A line was drawn between Pg' and N' and a line between Pg' and Ls, and the angle between the lines passing through these points was calculated as in Fig-17. To calculate the H angle in between, these points' x and y coordinate values were extracted as a 2- dimensional plane. With these coordinates, the slopes of these two lines were calculated using the formula in Fig- 18-b, and the angle H, which is the angle between the two lines, was calculated with the formula in Fig-18-a using these slopes.

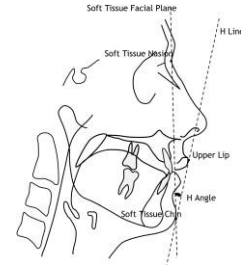


Figure 17. H-Angle of profile images.

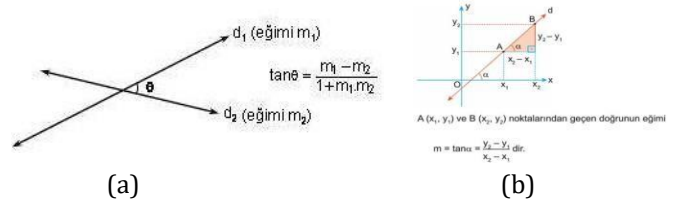


Figure 18. (a) Angle between 2 points, (b) slope of a line. The threshold value computed in section 3.2(rule based method) was incorporated as an additional feature. For TVL line feature, 68 reference points on the face were computed as x and y values by drawing a parallel line to the Sn (Subnasale) point and perpendicular to the transverse plane.

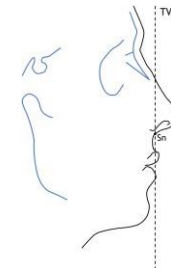


Figure 19. True Vertical Line (TVL).

The distances of each point to the TVL line were computed, and distance ratios were calculated as taking the TVL line as a reference. We aimed for a correct diagnosis from profile pictures taken by any mobile phone, but general variables of the images such as proximity, distance, location, and angle were not standard. To achieve higher accuracy with the nonstandard photos, we decided that our machine learning model should handle the lengths of the points from the TVL as they were scalable. For this purpose, we scaled all the profile photos by dividing the distances between the landmark points by a reference length. To pick the best reference length to scale by, we experimented with all paired combinations of 68 points. As a result, when the TVL was divided by the length between G and Gn, the most accurate distance to TVL scaling was obtained. The distances between the "Gn", "Sn", "A", "UL", "LL", "B" and "Pog" points on the face and the TVL were computed.

3.3.2. Machine Learning Model

Firstly, we divided our data set into 75% training and 25% testing. Next, we selected the most appropriate model and parameter optimization on 75% of the training data. We calculated the correlations for the features, evaluated the features' correlation against each other, and removed the "Pog" data, as its correlation was greater than 0.80. No data was removed since we created our dataset by feature extraction from images.

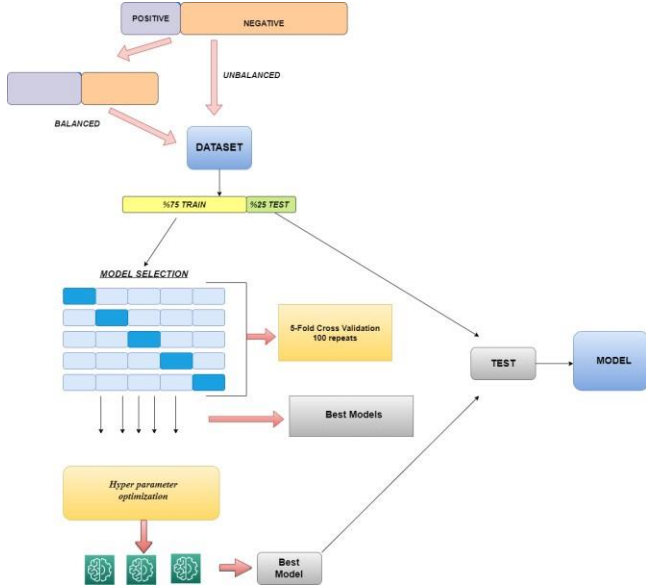


Figure 20. Experimental set-up for Machine Learning.

Since the data set, we used in the machine learning method was not balanced, we created a balanced data set with random selection of the negative class and kept the entire positive class. We compared the results by repeating the experiment twice for our balanced (Table-6) and unbalanced data sets (Table-7). In this experiment, we used 5-fold cross-validation for model selection and repeated this experiment 100 times on 18 models (Fig-20). We selected the model with the best results after the cross-validation. After hyperparameter optimization, the best parameters were selected. By creating our models according to these parameters, we tested them on the test data and selected the model with the best results.

As the Random Forest, Logistic regression, Ridge Classifier, Extra Trees Classifier, Linear SVM performed best, and we optimized the parameters on these five models to select the most optimal parameters. In addition, we applied parameter optimization on a few more models with close cross validation performances.

Table 7. Balanced dataset results

	Accuracy	Precision	Recall	F1	AUC
LR	0.759 (0.071)	0.753 (0.084)	0.781 (0.115)	0.761 (0.076)	0.841 (0.065)
RC	0.765 (0.069)	0.754 (0.080)	0.796 (0.101)	0.770 (0.069)	0.842 (0.065)
SGD	0.681 (0.094)	0.739 (0.150)	0.684 (0.260)	0.658 (0.144)	0.836 (0.073)

PAC	0.604 (0.097)	0.689 (0.200)	0.692 (0.335)	0.595 (0.171)	0.822 (0.073)
KNNC	0.742 (0.068)	0.729 (0.076)	0.779 (0.103)	0.748 (0.069)	0.800 (0.069)
DT	0.656 (0.075)	0.661 (0.092)	0.641 (0.120)	0.644 (0.087)	0.653 (0.076)
ETC	0.653 (0.082)	0.660 (0.096)	0.652 (0.123)	0.651 (0.088)	0.652 (0.084)
L-SVC	0.754 (0.070)	0.754 (0.085)	0.764 (0.126)	0.752 (0.077)	0.837 (0.065)
SVC	0.765 (0.071)	0.759 (0.085)	0.785 (0.104)	0.767 (0.073)	0.836 (0.067)
NB	0.732 (0.069)	0.795 (0.108)	0.645 (0.128)	0.701 (0.086)	0.802 (0.074)
ABC	0.682 (0.075)	0.685 (0.086)	0.683 (0.117)	0.678 (0.081)	0.730 (0.081)
BC	0.722 (0.071)	0.743 (0.093)	0.685 (0.117)	0.703 (0.087)	0.783 (0.076)
RF	0.773 (0.070)	0.788 (0.090)	0.760 (0.115)	0.765 (0.079)	0.829 (0.069)
ETSC	0.773 (0.069)	0.788 (0.085)	0.751 (0.105)	0.765 (0.074)	0.824 (0.069)
GPC	0.769 (0.070)	0.736 (0.075)	0.847 (0.091)	0.784 (0.066)	0.815 (0.072)
GBC	0.728 (0.070)	0.735 (0.085)	0.726 (0.111)	0.723 (0.075)	0.790 (0.072)
LDA	0.730 (0.072)	0.736 (0.084)	0.726 (0.112)	0.725 (0.078)	0.815 (0.070)
QDA	0.657 (0.068)	0.757 (0.125)	0.482 (0.151)	0.571 (0.115)	0.756 (0.076)

Table 8. Unbalanced dataset results.

	Accuracy	Precision	Recall	F1	AUC
LR	0.825 (0.037)	0.722 (0.114)	0.510 (0.106)	0.591 (0.095)	0.859 (0.047)
RC	0.816 (0.032)	0.747 (0.125)	0.423 (0.106)	0.531 (0.099)	0.859 (0.047)
SGD	0.776 (0.100)	0.638 (0.175)	0.581 (0.224)	0.570 (0.106)	0.852 (0.050)
PA	0.661 (0.220)	0.579 (0.299)	0.455 (0.362)	0.372 (0.189)	0.845 (0.056)
KN	0.794 (0.042)	0.616 (0.108)	0.519 (0.117)	0.556 (0.095)	0.772 (0.060)
DT	0.733 (0.049)	0.473 (0.102)	0.491 (0.123)	0.475 (0.098)	0.654 (0.065)
ETC	0.725 (0.051)	0.458 (0.104)	0.469 (0.126)	0.451 (0.104)	0.638 (0.063)
L-SVC	0.824 (0.035)	0.733 (0.117)	0.483 (0.108)	0.577 (0.093)	0.858 (0.047)
SVC	0.819 (0.033)	0.729 (0.121)	0.466 (0.108)	0.560 (0.096)	0.853 (0.046)
NB	0.794 (0.047)	0.629 (0.126)	0.517 (0.134)	0.553 (0.097)	0.810 (0.059)
ABC	0.772 (0.041)	0.560 (0.103)	0.480 (0.119)	0.510 (0.097)	0.768 (0.068)
BC	0.779 (0.042)	0.597 (0.116)	0.441 (0.113)	0.499 (0.098)	0.784 (0.057)
RF	0.794 (0.037)	0.624 (0.102)	0.471 (0.114)	0.525 (0.097)	0.814 (0.054)
ETSC	0.792 (0.038)	0.617 (0.107)	0.467 (0.117)	0.528 (0.095)	0.802 (0.054)
GPC	0.810 (0.039)	0.680 (0.122)	0.485 (0.108)	0.558 (0.097)	0.819 (0.054)
GB	0.788 (0.041)	0.606 (0.110)	0.482 (0.116)	0.530 (0.098)	0.805 (0.055)
LD	0.814 (0.040)	0.689 (0.121)	0.499 (0.109)	0.571 (0.098)	0.842 (0.050)
QDA	0.759 (0.036)	0.560 (0.201)	0.253 (0.151)	0.326 (0.145)	0.783 (0.062)

3.3.3. Hyper parameter optimization

We made hyperparameter optimization over the selecting models.

Table 9. Hyperparameter optimization results for four models.

Random Forest	Gradient Boosting
'max_depth': 20	'max_depth': 20
'max_features': 'auto'	'max_features': 'auto'
'min_samples_leaf': 1	'min_samples_leaf': 1
'n_estimators': 200	'n_estimators': 200
Logistic Regression	Ridge Classifier
'C': 10	'alpha': 0.1
'penalty': 'l2'	
'solver': 'newton-cg'	

3.3.4. Test Accuracies

Although the unbalanced dataset resulted in higher accuracy, F1 and precision were low. Therefore, balanced dataset's accuracy, F1 and recall score more represent the actual production accuracy. The best result for the balanced machine learning model was Logistic Regression with 76% accuracy, 74% precision and 77% F1, 82% recall, and AUC of 76%.

Table 10: Balanced data test results for each model.

Random Forest Classifier	
Accuracy	0.672
Precision	0.666
F1	0.689
Recall	0.714
AUC	0.671
Gradient Boosting Classifier	
Accuracy	0.763
Precision	0.758
F1	0.771
Recall	0.785
AUC	0.763
Ridge Classifier	
Accuracy	0.763
Precision	0.741
F1	0.779
Recall	0.821
AUC	0.762
Linear SVC	
Accuracy	0.763
Precision	0.758
F1	0.771
Recall	0.785
AUC	0.763
Logistic Regression	
Accuracy	0.763
Precision	0.741
F1	0.779
Recall	0.821
AUC	0.762

Table 11: Unbalanced data test results for each model

Logistic Regression	
Accuracy	0.798
Precision	0.625
F1	0.576
Recall	0.535
AUC	0.712
Random Forest Classifier	
Accuracy	0.798
Precision	0.636
F1	0.560
Recall	0.500
AUC	0.700
Gradient Boosting Classifier	
Accuracy	0.825
Precision	0.680
F1	0.641
Recall	0.607
AUC	0.754
Ridge Classifier	
Accuracy	0.834
Precision	0.777
F1	0.608
Recall	0.500
AUC	0.725
Linear SVC	
Accuracy	0.816
Precision	0.681
F1	0.600
Recall	0.535
AUC	0.724
Extra Trees Classifier	
Accuracy	0.834
Precision	0.750
F1	0.625
Recall	0.535
AUC	0.736

4. Conclusion

This study presents the first fully automated machine learning method to accurately diagnose Class-III malocclusion applied across mobile images, to the best of our knowledge. For this purpose, we comparatively evaluated three machine learning approaches. We achieved 70% accuracy with the deep learning method, 74% accuracy with the area ratio method and 76% success with the machine learning method.

All three of our methods are flexible to be adapted to any dynamic training set of profile images of any ethnicity and age. Our next goal is to integrate our most accurate learning model among these three into a mobile application for parents and pediatricians seeking a second opinion on whether to reach out to an orthodontist at an early stage of developmental bone growth with a warning of Class-III malocclusion risk.

5. Acknowledgment

We want to thank Gül Sude Demircan, who developed the previous prototype, and Tülay Sevinç, who assisted in collecting the patient images and the consent forms.

6. References

- [1] Staudt CB, Kiliaridis S. "Different skeletal types underlying Class-III malocclusion in a random population." *Am J Orthod Dentofacial Orthop*, 136(5), 715-721, 2009.
- [2] Oltramari-Navarro PV, de Almeida RR, Conti AC, Navarro Rde L, de Almeida MR, Fernandes LS. "Early treatment protocol for skeletal Class-III malocclusion." *Braz Dent J*, 24(2), 167-173, 2013.
- [3] Al-Khalifa, Hussein. (2014). "Orthopedic Correction of Class-III Malocclusions during Mixed Dentition." *Open Journal of Stomatology*. 04(07), 372-380,2014
- [4] Mandall N, Cousley R, DiBiase A, Dyer F, Littlewood S, Mattick R, Nute SJ, Doherty B, Stivaros N, McDowall R, Shargill I, Worthington HV. "Early Class-III protraction facemask treatment reduces the need for orthognathic surgery: a multi-centre, two-arm parallel randomized, controlled trial." *J Orthod.*, 43(3), 164-175, 2016.
- [5] Sharma JN. "Epidemiology of malocclusions and assessment of orthodontic treatment need for the population of eastern Nepal." *World J Orthod.*, 10(4), 311- 316, 2009.
- [6] X. Xu et al., "Advances in Smartphone-Based Point-of-Care Diagnostics," in *Proceedings of the IEEE*, vol. 103, no. 2, pp. 236-247, Feb. 2015, doi: 10.1109/JPROC.2014.2378776.
- [7] Digital around the world - datareportal – global digital insights. *DataReportal*. (n.d.). Retrieved July 25, 2022, from <https://datareportal.com/global-digital-overview>
- [8] Mobile Health Industry Trends and forecast 2021. *Artezio*. (n.d.). Retrieved July 24, 2022, from <https://www.artezio.com/pressroom/blog/mobile-industry-forecast/>
- [9] Gupta G, Vaid NR. "The World of Orthodontic apps." *APOS Trends Orthod*, 7(2), 73, 2017.
- [10] Development, C. S. (n.d.). *Dental4Windows*. *Download.com*. Retrieved July 24, 2022, from

https://download.cnet.com/Dental4Windows/3000-2129_4-76472046.html

- [11] Baheti, M.J., Toshniwal, N. "Orthodontic apps at fingertips.", *Progress in Orthodontic*, 15(1), 36, 2014.
- [12] Phimentum. (n.d.). Retrieved July 23, 2022, from <https://www.phimentum.com/>
- [13] Demircan, G.S., Kılıç, B., Önal-Süzek, T. (2021). "Early Diagnosis and Prediction of Skeletal Class-III Malocclusion from Profile Photos Using Artificial Intelligence." In: Jarm, T., Cvetkoska, A., Mahnič-Kalamiza, S., Miklavcic, D. (eds) 8th European Medical and Biological Engineering Conference. *EMBEC 2020. IFMBE Proceedings*, 80, 434-448, 2020.
- [14] Basciftci,F.A.,Uysal,T.,Buyukerkmen,A."Determinati on of Holdaway soft tissue norms in Anatolian Turkish adults" *Am J Orthod Dentofacial Orthop*, 123(4),395-400, 2003.
- [15] 1adrianb. (n.d.). 1adrianb/face-alignment: 2D and 3D face alignment library build using pytorch. *GitHub*. Retrieved July 24, 2022, from <https://github.com/1adrianb/face-alignment>.

Appendix

LR	Logistic Regression
RC	Ridge Classifier
SGD	SGD Classifier
PA	Passive Aggressive Classifier
KN	K Neighbors Classifier
DT	Decision Tree Classifier
ETC	Extra Tree Classifier
L-SVC	Linear SVC
SVC	SVC
NB	Gaussian NB
ABC	Ada Boost Classifier
BC	Bagging Classifier
RF	Random Forest Classifier
ETSC	Extra Trees Classifier
GPC	Gaussian Process Classifier
GB	Gradient Boosting Classifier
LD	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis