scientific reports

OPEN

Reproducibility analysis of automated deep learning based localisation of mandibular canals on a temporal CBCT dataset

Jorma Järnstedt[1,2 ✉], Jaakko Sahlsten[3], Joel Jaskari[3], Kimmo Kaski[3,6 ✉], Helena Mehtonen[1], Ari Hietanen[4], Osku Sundqvist[4], Vesa Varjonen[4], Vesa Mattila[4], Sangsom Prapayasatok[5] & Sakarat Nalampang[5]

Preoperative radiological identification of mandibular canals is essential for maxillofacial surgery. This study demonstrates the reproducibility of a deep learning system (DLS) by evaluating its localisation performance on 165 heterogeneous cone beam computed tomography (CBCT) scans from 72 patients in comparison to an experienced radiologist's annotations. We evaluated the performance of the DLS using the symmetric mean curve distance (SMCD), the average symmetric surface distance (ASSD), and the Dice similarity coefficient (DSC). The reproducibility of the SMCD was assessed using the within-subject coefficient of repeatability (RC). Three other experts rated the diagnostic validity twice using a 0–4 Likert scale. The reproducibility of the Likert scoring was assessed using the repeatability measure (RM). The RC of SMCD was 0.969 mm, the median (interquartile range) SMCD and ASSD were 0.643 (0.186) mm and 0.351 (0.135) mm, respectively, and the mean (standard deviation) DSC was 0.548 (0.138). The DLS performance was most affected by postoperative changes. The RM of the Likert scoring was 0.923 for the radiologist and 0.877 for the DLS. The mean (standard deviation) Likert score was 3.94 (0.27) for the radiologist and 3.84 (0.65) for the DLS. The DLS demonstrated proficient qualitative and quantitative reproducibility, temporal generalisability, and clinical validity.

The identification of bilateral mandibular canals from preoperative radiological examinations is important before oral and maxillofacial surgical operations in the lower jaw area, such as implantology, impacted teeth operative extraction, and pathological lesion resection or enucleation[1]. The mandibular canals are seen as radiolucent structures surrounded by cortical bony borders with two openings; the foramen mentale in the anterior and foramen mandibulae in the posterior part. The thickness, location, and shape of the mandibular canal can have a number of anatomical variations, and some of them are impossible to be identified due to difficult bone quality and pathological changes[2,3]. These canals host a neurovascular bundle that includes an artery, veins, and the inferior alveolar nerve that supplies motoric and sensory innervations to the chin, lower lip, and teeth. Damaging the mandibular canals may cause excessive bleeding, and temporary or permanent neurological complications[4]. The challenging and labour intensive localisation of these canals is currently done manually, which poses a challenge, as there are not enough specialists to meet the diagnostic needs[5]. Automated tools show promise to alleviate the manual burden[6,7].

In order to locate the mandibular canals accurately, three dimensional (3D) radiological imaging is the recommended approach. For this purpose, Cone Beam Computed Tomography (CBCT) is commonly used[8]. However, there are diagnostic challenges due to technical and patient related artefacts, as well as heterogeneities. In 2020, the International Society for Strategic Studies in Radiology published recommendations for technical validation in radiological AI research with four crucial properties: accuracy, robustness, generalisability, and reproducibility. The accuracy provides insight to the expected performance, robustness to the invariance to small perturbations

[1]Medical Imaging Centre, Department of Radiology Tampere University Hospital, Teiskontie 35, 33520 Tampere, Finland . [2]The Graduate School, Chiang Mai University, 239 Huaykaew Road, Suthep, Mueang, Chiang Mai, Thailand . [3]Aalto University School of Science, Otakaari 1, 02150 Espoo, Finland . [4]Planmeca Oy, Asentajankatu 6, 00880 Helsinki, Finland . [5]Division of Oral and Maxillofacial Radiology, Faculty of Dentistry, Chiang Mai University, Suthep Rd., T. Suthep, A. Muang, Chiang Mai, Thailand . [6]Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK. ✉email: jorma.jarnstedt@pirha.fi; kimmo.kaski@aalto.fi

in the data, and generalisability to the behaviour with various medical imaging variabilities[9]. The reproducibility is defined as invariance to measurement variability on the same subject under changing conditions, which can be evaluated from time-wise separated scans[10]. There are several reasons for scanning patients multiple times, such as pathological changes, operation planning, and treatment follow-up[11]. Without evaluating reproducibility, there is no evidence that a same image derived feature or element could be used in a longitudinal assessment[9].

Imaging-based deep learning models perform computations on the voxels of a scan. When a patient is scanned at different times, there are always at least slight changes in the voxel intensities, even when the patient is scanned again using the same device and imaging parameters. These changes alter the internal computations of the deep learning models, but the effect on the model output should be negligible in the case of a robust and reproducible model. Therefore, there is a close relationship between the reproducibility and the overall performance of the model[12].

Recently, multiple studies have introduced deep learning based automatic mandibular canal segmentation systems that have shown impressive accuracy and robustness[13–16]. Two of these studies also reported the effect of heterogeneities on the performance of deep learning systems[13,16], and one of them the deep learning generalisability and interobserver variability[16]. However, none of these works have analysed the reproducibility of the deep learning systems.

The aim of this study is to validate the reproducibility and temporal generalisation of a previously introduced deep learning based automatic mandibular canal localisation system[13,16], both quantitatively and qualitatively. Our study contributes substantially to the field by demonstrating the reproducibility of the deep learning system that is required for follow-up clinical validation of the system and its use in longitudinal studies.

Methods

In this section, we present a detailed description of the deep learning system and its components, describe the patient data used for the results, and the quantitative and qualitative evaluation measures we use to evaluate reproducibility.

Deep learning model. The deep learning system (DLS) we study was proposed in two recent studies[13,16]. This model was also used as a basis in the work of publicly available voxel level annotated 3D mandibular canal dataset[14]. The DLS consists of two algorithms; a deep learning neural network that segments the voxels containing the mandibular canals from 3D CBCT volumes, and a postprocessing algorithm that extracts the two most likely mandibular canal curves from the segmentation volume. For our analysis, we utilise the previously trained model[16], which was developed on a heterogeneous, multi-device, and multiethnic dataset of 1082 scans to evaluate its temporal generalisability and reproducibility with a new dataset, focused on challenging clinical heterogeneities.

The deep learning model of DLS is a type of fully convolutional neural network with a U-net style architecture[17]. The model utilises three dimensional convolutional layers that enable the recognition of patterns simultaneously in the axial, coronal, and sagittal planes. The deep learning model produces approximate segmentations of the mandibular canals, which are then postprocessed using connected component analysis for the purpose of reducing false positives as well as connecting partial segmentations to form a single canal. Segmentation is then reduced into a curve using a skeletonisation routine[18], and a heuristic concatenation algorithm is used to select and combine curve segments in order to extract the two mandibular canals based on correct anatomical characteristics and symmetry.

Patient data. The data was collected retrospectively from the Tampere University Hospital (TAUH) in Tampere, Finland. The patients whose data had been previously used in the development, validation, or testing of the DLS were excluded. Every patient has at least two CBCT scans imaged for medical reasons. These reasons include benign lesions follow-up, operation planning and control for orthognathic surgery and temporomandibular joint prosthesis, trauma controls, computer assisted surgery planning, and bone destructive lesions including malignancies. The collected dataset of 165 CBCT scans consisted of 72 patients with a mean age of 45.2±17.0 and a range from 17 to 89 years, having 47% males with a mean age of 43.7±15.8 and a range from 20 to 72 years, and 53% females with a mean age of 46.7±18.3 and a range from 17 to 89 years. Patients' age information was collected according to age at the last repeated scan. The scans were pseudonymised and categorised into four main subgroups to evaluate the effects of different heterogeneities. In addition, the dataset included immediate postoperative scans that were categorised into a separate heterogeneity subgroup, to examine the effect of the surgical operation before the healing process.

Three scanners with different imaging parameters were used; KAVO OP 3D PRO, VISO G7, and Scanora 3Dx with 7, 66, and 92 scans, respectively. The voxel spacing of the scans ranged from 0.2 mm to 0.5 mm with the most common ones being 0.2 mm, 0.3 mm, and 0.4 mm with 57%, 29%, and 7% of the scans, respectively. The dataset consisted of 107 high resolution (HR), 33 standard resolution (SR), 12 low dose (LD), and 13 ultra-low dose (ULD) scans. The distribution of different devices, voxel spacings, and doses are presented in Supplementary Table S1. All the scans were whole facial imaging, large field-of-view scans, due to the aforementioned diagnostic needs in radiological imaging. There were 39 patients with two CBCT scans, and 33 patients with multiple (three to five) scans in the study material. There were 11 patients who were scanned with the same CBCT device and the same imaging parameters, 16 scanned with the same device but different imaging parameters, and 45 who were scanned with different devices and different imaging parameters. All the scans were taken in TAUH from 2015 to 2020.

The mandibular canals were annotated using clinical software: Romexis 4.6.2, of Planmeca, Finland, by a senior radiologist with over 35 years of experience in dentistry, similarly to previous works[13,16]. In the annotation

process, a CBCT panoramic reconstruction was made (Fig. 1a), cross-sectional images against panoramic curves were reconstructed (Fig. 1b), and the midline of the canals were marked with standard 1.5 mm diameter control points placed in 3 mm intervals. The whole path of the mandibular canal was marked starting from the foramen mentale opening and ending at the foramen mandibulae opening. Multiple control points were used for each of the canals, especially in the foramen mentale curvature area and in complicated cases (Fig. 1c). An approximate segmentation was created when computing segmentation measures, in which the path of the canal was expanded to a 1.5 mm static diameter tube.

The four main subgroups manifest various clinical and operative situations in dental and maxillofacial radiology, including pre- and postoperative, and follow-up CBCT scans. These subgroups were determined as *Normal group*, *Prosthetic group*, *Orthognathic surgery group*, and *Pathological group* with 23, 13, 22, 14 patients and 51, 35, 46, 33 scans, respectively.

The patients of Normal group were scanned for reasons having no major influence in the mandibular canal area, but 14 of the 51 scans (27%) were reported to have difficult bone structure. The patients of Prosthetic group have major temporomandibular joint (TMJ) disorders and have at least one preoperative scan and one postoperative scan after full prosthetic metallic TMJ-reconstruction. An example of a patient from TMJ Prosthetic group is shown in (Fig. 2). The patients of Orthognathic surgery group have undergone a bilateral sagittal split osteotomy (BSSO), where the mandible is split bilaterally and moved forward or backward, to correct malocclusion and functionality[19], and have at least one preoperative and one postoperative scan taken six to 12 months after the operation. The BSSO changes the path of the mandibular canal and fixation materials cause metallic artefacts in the imaged area, both seen as differences between the preoperative scan and postoperative scan of the patient. The effect of this heterogeneity is shown in three dimensional (3D) reconstructions models of preoperative (Fig. 3a) and postoperative (Fig. 3b) cases. The patients of Pathological group have severe pathological bone destructive diseases in the mandible, which may partly or completely destroy the visibility of the mandibular canals, and have at least one preoperative, and one postoperative or follow-up scan. Eight of the 14 patients (57%) had a malignant bone destruction on the right side of the mandible.

The immediate postoperative scans, i.e. those imaged less than five days after the BSSO surgery, were excluded from the main analysis. The visibility of mandibular canal is extremely poor in these scans[20] and there are changes to the path of the mandibular canal due to BSSO. Thus, the localisation of the mandibular canal is challenging in these scans. However, the task of localising the mandibular canal in these cases is rarely clinically relevant and thus the analysis is included in the Supplementary.

This study is based on a retrospective and registry-based dataset, and as such does not involve experiments on humans and/or the use of human tissue samples and no patients were imaged for this study. A registration and retrospective study does not need ethical permission or informed consent from subjects according to the law of Finland (Medical Research Act (488/1999) and Act on Secondary Use of Health and Social Data (552/2019))
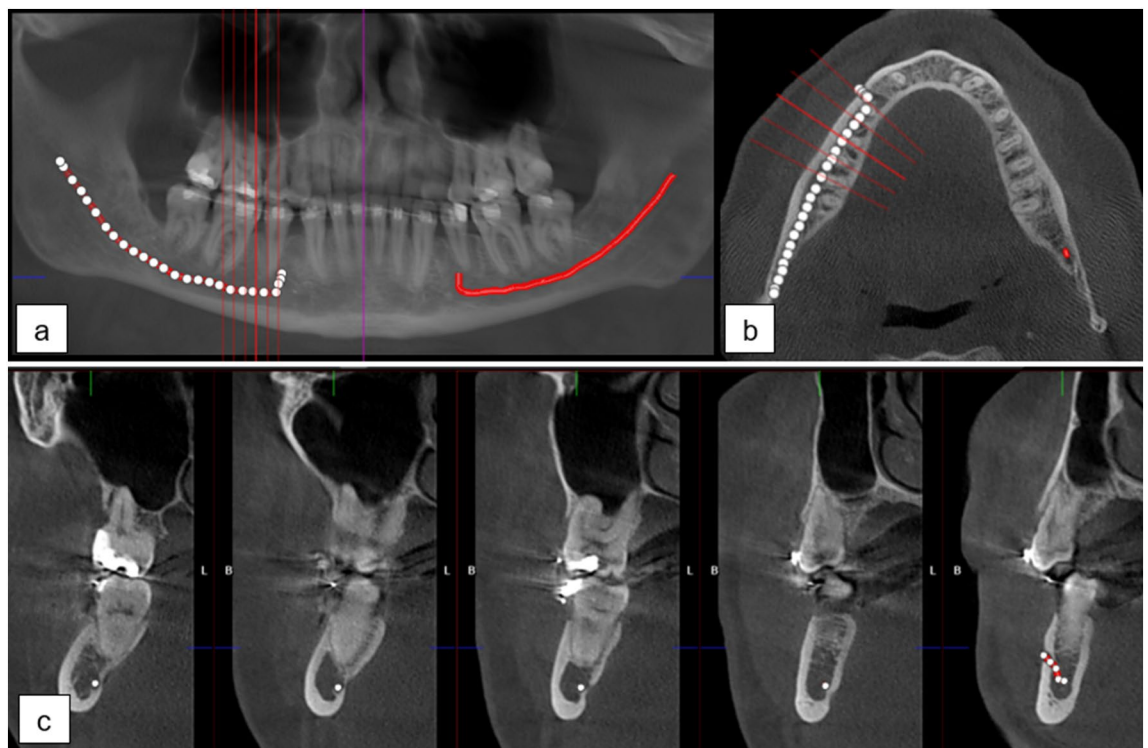

Figure 1. (**a**) CBCT panoramic reconstruction showing the annotation control points in the left and the interpolated mandibular canal in right side of the mandible, (**b**) an axial slice with cross-sectional slices along the panoramic curve shown in red, (**c**) annotation control points in the cross-sectional slices 3 mm apart with multiple annotation points in the foramen mentale area.
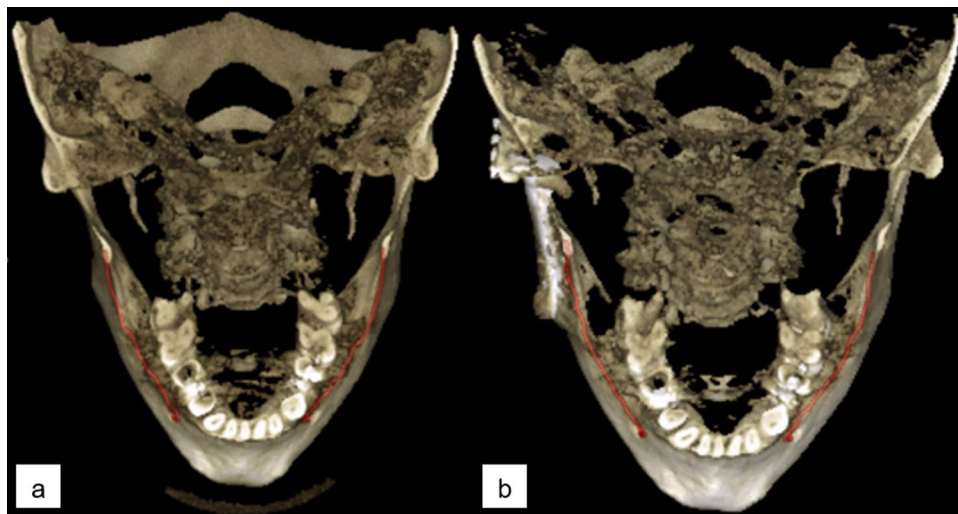
Figure 2. Three dimensional reconstruction images with manually annotated mandibular canals shown in red, (**a**) preoperative CBCT scan, (**b**) postoperative CBCT scan with right side temporomandibular joint reconstruction on the angulus area of the mandible.
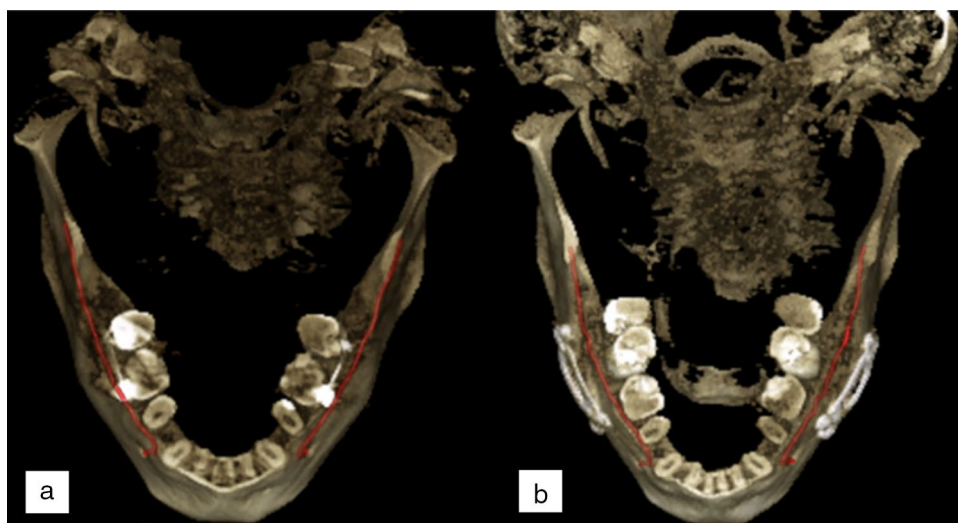

Figure 3. Three dimensional reconstruction images of an orthognathic surgery patient showing changes in the postoperative scan with manually annotated mandibular canals shown in red, (**a**) preoperative CBCT scan, (**b**) postoperative CBCT scan 1 year after the operation.

and according to European General Data Protection Regulation (GDPR) rules 216/679. The use of the data was accepted by the Tampere University Hospital Research Director, Tampere, Finland, 1st of October, 2019 (vote number R20558).

Quantitative analysis approach. In the quantitative analysis, the outputs of the DLS were compared to the senior radiologist's annotations. The mandibular canal localisation accuracy of the DLS was evaluated using the Symmetric Mean Curve Distance (SMCD)[16]. In the SMCD, the average deviation between the full path of the radiologist's annotation and the DLS output is determined. It thus measures how far apart the two curves are on average. In addition, the segmentation performance of the DLS was evaluated using the Dice similarity coefficient (DSC), which is a normalized measure of the intersection of two segmentations, and the average symmetric surface distance (ASSD) that measures the average error between the surfaces of two segmentations. However, these metrics are computed using the senior radiologist's approximate segmentation of the canal as the ground truth, and thus they only approximate the true segmentation performance.

The reproducibility was evaluated by measuring the temporal variability in SMCD for each canal, i.e. the two canals of each patient were treated as independent subjects of interest. The reproducibility measures we used were the within-subject standard deviation (wSD) and the repeatability coefficient (RC), and we also report the

within-subject mean and ranges of values. The wSD estimates the variability in SMCD values for the canals and the RC the largest difference between two SMCD values for the same canal with 95% confidence. These quantities are computed in Eqs. (1) and (2) as follows[21]:

$$wSD = \sqrt{WMS}, \tag{1}$$

$$RC = 1.96\sqrt{2}\, wSD$$
$$\approx 2.77 wSD, \tag{2}$$

where WMS is the within-subject mean squared error, computed as:

$$WMS = \sum_{i=1}^{n} \sum_{k=1}^{K_i} \frac{(Y_{ik} - \bar{Y}_i)^2}{n(K-1)},$$

n is the number subjects, K_i the number of repetitions for subject i, Y_{ik} the kth measurement of subject i, and $\bar{Y}_i$ the mean of replications of subject i. As our dataset includes variable number of repetitions per subject, the K is computed by applying a downward correction to reduce overestimation of the variation among smaller groups compared to larger groups: $K = \frac{1}{n-1}\left(N - \frac{\sum_{i=1}^{n} K_i^2}{N}\right)$, where N is the total number of scans[22].

Qualitative analysis approach. For the qualitative analysis, the annotations by the senior radiologist and those produced by the DLS were assessed by three experts; two dental and maxillofacial radiologists with 15 and 13 years of working experience, and one resident of dental and maxillofacial radiology with 8 years of working experience in dentistry.

There were in total 636 mandibular canal segmentations in the comparison: 318 annotated by the radiologist and 318 DLS outputs, which were provided to the three experts independently and in a random order, without informing whether the segmentation was done by the radiologist or by the DLS. In addition, the segmentations were assessed two times by each expert in separate sessions with a two week interval. The observers assessed the quality of each segmentation and its feasibility for clinical diagnostics using a five-point Likert scale defined with 0: not usable for diagnostics, 1: partly usable for diagnostics, canal visibility below one half, 2: minor issues with almost fully usable for diagnostics, 3: almost perfect, fully diagnostically usable, and 4: perfect, fully diagnostically usable. The Likert ratings were also used to form a binary scale for the diagnostic suitability of a canal segmentation. It was defined as Likert ratings 3 and 4 being fully suitable for diagnosis and 0–2 as not fully suitable.

The observers justified their ratings using five pre-selected error types: fully missing, major parts of the segmentation missing, slightly off the centre of the canal, short at the mandibular foramen, and short at the mental foramen. The first and second error types are considered clinically relevant resulting in Likert rating of 0-2 and diagnostically not fully usable. The other three error types have less clinical importance and they can be explained by the anatomy of the mandible, and the nature of the heterogeneity subgroups.

The reproducibility was evaluated using the Likert scores of the first annotation session. Due to the discrete and ordinal nature of the Likert scores, as well as the high consistency of the given scores in the dataset, Gaussian random effects models and analysis of variance methods were deemed unsuitable. In addition, the number of temporal scans varied between the patients, and thus methods that require paired data, such as the Kendall rank correlation coefficient, could not be used to analyse the reproducibility. We utilised a Bayesian statistical analysis approach for the reproducibility of ordinal measurements, proposed in a recent study[23]. In short, it uses a random effects model, with subject-dependent random effects, combined with the Master's partial credit model[24]. It defines the probability that the subject i's Likert score y_i is equal to the number h, given the random effect of the subject x_i and grader j specific parameters α_j and δ_j, as follows:

$$p_j(y_i = h \mid x_i, \alpha_j, \delta_j) = \frac{\exp\left(\sum_{m=1}^{h-1} \alpha_j (x_i - \delta_{jm})\right)}{\sum_{n=1}^{H} \exp\left(\sum_{m=1}^{n-1} \alpha_j (x_i - \delta_{jm})\right)}.$$

It should be noted that in computing the probabilities, the Likert scores were mapped from 0–4 to 1–5. A detailed description of the approach is given in the Supplementary material.

We used the *repeatability measure* (RM) proposed in the aforementioned study to measure the reproducibility. For grader j, it is defined as the probability that two Likert scores for the same segmentation will be the same, averaged over the data in Eq. (3):

$$\hat{RM}_j(\alpha_j, \delta_j) = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{5} p_j(y_i = h \mid x_i, \alpha_j, \delta_j)^2 \tag{3}$$

The Bayesian analysis produces a distribution of possible parameter values α_j and δ_j, which we used to compute the posterior average $RM_j = Ave_{\alpha_j, \delta_j}(\hat{RM}_j)$ and the 95% credibility intervals for each Expert. The analysis was performed for the full dataset, as well as for each heterogeneity group independently. We implemented the analysis using the Stan[25] programming language with the Python interface PyStan[26], and using the default settings described in the original study[23].

To analyse the intra- and interobserver variability, we calculated the accuracy of Likert ratings. The interobserver variability was evaluated by comparing the assessments between each pair of experts, and the intraobserver variability by comparing the ratings of the first and second assessment session.

Results

This section presents the results for the quantitative analysis of the DLS performance, and for the qualitative analysis approach for the radiologist's annotations and the DLS performance.

Quantitative analysis. The DLS had a repeatability coefficient of 0.969 mm for the full dataset, and when evaluated for each heterogeneity group, the RC was 0.329 mm for Normal, 0.574 mm for TMJ Prosthetic, 1.707 mm for Orthognathic, and 0.648 mm for Pathological group. Full results are shown in Table 1. The overall localisation performance of the DLS, measured using the median and interquartile range (IQR) of SMCD on valid canals, was found to be 0.643 (0.186) mm for the full dataset, and 0.638 (0.151) mm, 0.683 (0.197) mm, 0.617 (0.178) mm, and 0.639 (0.201) mm for Normal, TMJ Prosthetic, Orthognathic, and Pathological heterogeneity groups, respectively. The segmentation performance of the DLS measured using the median and IQR of average symmetric surface distance (ASSD) on valid canals was 0.351 (0.135) mm for the full dataset, and 0.328 (0.088) mm, 0.356 (0.133) mm, 0.365 (0.148) mm, and 0.355 (0.145) mm for Normal, TMJ Prosthetic, Orthognathic, and Pathological heterogeneity groups, respectively. The Dice similarity coefficient (DSC) on all the canals was 0.567 (0.133) for the full dataset, and 0.597 (0.121), 0.550 (0.114), 0.540 (0.163), and 0.579 (0.108) for Normal, TMJ Prosthetic, Orthognathic, and Pathological heterogeneity groups, respectively. Full results of the localisation and segmentation performance of the DLS for each of the heterogeneity groups are shown in Fig. 4. In addition, the patient level reproducibility is presented in Fig. 5. Moreover, analysis of the overall quantitative DLS performance when grouped by age, gender, device, and dose as well as quantitative reproducibility analysis of DLS and radiologists, when grouped by age, gender, and device are reported in the Supplementary.

Qualitative analysis. In the qualitative analysis, the reproducibility of the Likert rating for the radiologist and the DLS, measured in the repeatability measure, was found to be 0.958 and 0.895 for Normal, 0.841 and 0.887 for TMJ Prosthetic, 0.945 and 0.913 for Orthognathic, and 0.916 and 0.886 for Pathological groups, respectively. In terms of diagnostically fully usable canal paths, the RM values for the radiologist and the DLS were 0.992 and 0.954 for Normal, 0.963 and 0.974 for TMJ Prosthetic, 0.978 and 0.922 for Orthognathic, and 0.997 and 0.980 for Pathological groups, respectively. Full results are shown in Table 2. Additional qualitative overall performance and reproducibility analysis when grouping by age, gender, and device are reported in the Supplementary.

Heterogeneity	n	N	K	Mean	wSD	RC	Range
Full dataset	131	302	{2, 3, 4, 5}	0.761	0.350	0.969	[0.471, 3.014]
Normal	44	98	{2, 3, 4}	0.743	0.119	0.329	[0.554, 3.014]
TMJ Prosthetic	26	70	{2, 3, 4, 5}	0.756	0.207	0.574	[0.487, 1.172]
Orthognathic	40	84	{2, 4}	0.806	0.616	1.707	[0.471, 2.206]
Pathological	21	50	{2, 3, 4}	0.719	0.234	0.648	[0.515, 1.049]

Table 1. Within-subject mean and standard deviation (wSD), repeatability coefficient (RC), and the range of valid SMCD (mm) values. The results are presented for the full dataset and for each heterogeneity group separately. *N* The total number of canals, *n* The number of unique canals, and *K* The set of the number of canals present in the group.
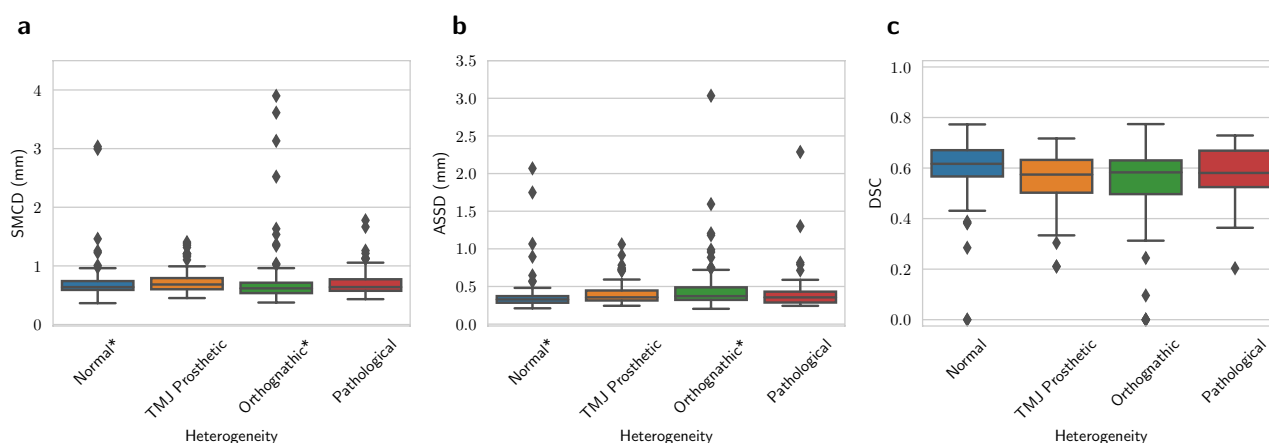

Figure 4. Quantitative performance of the DLS for each heterogeneity group. (**a**) Symmetric Mean Curve Distance (mm), (**b**) Average symmetric surface distance (mm), (**c**) Dice similarity coefficient. *Two and four values omitted in distance measures from Normal and Orthognathic groups, respectively, as the DLS did not find the canals for these cases.
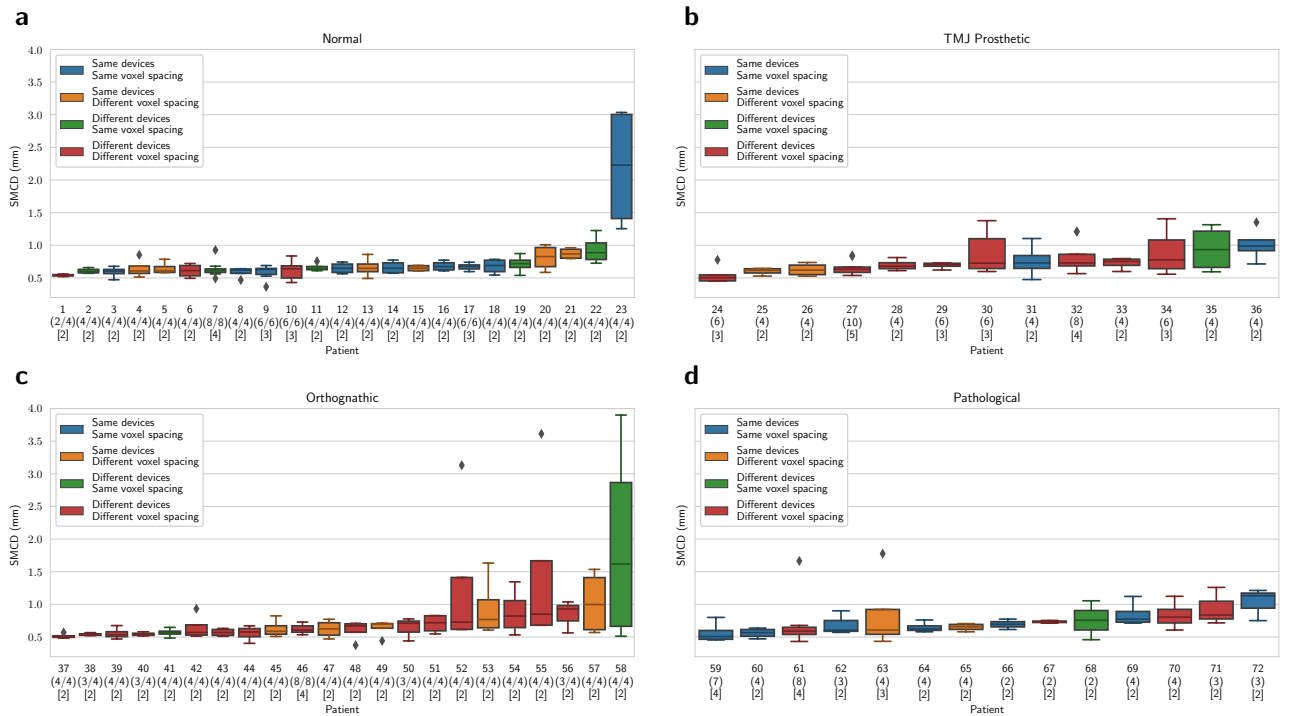
Figure 5. DLS performance evaluated with SMCD for each patient. (**a**) Normal, (**b**) TMJ Prosthetic, (**c**) Orthognathic, and (**d**) Pathological heterogeneity group. Parentheses show the number of valid canals out of the total number of canals, and the square brackets the number of scans for each patient. Each patient is colored with the temporal imaging configuration, indicating whether the scans are from the same or different devices, and with the same or different voxel spacings.

Reproducibility	Heterogeneity	Radiologist mean (95% CI)	DLS mean (95% CI)
Average	Full dataset	0.923 (0.842, 0.979)	0.877 (0.814, 0.932)
	Normal	0.958 (0.882, 1.000)	0.895 (0.833, 0.935)
	TMJ Prosthetic	0.841 (0.725, 0.925)	0.911 (0.829, 0.970)
	Orthognathic	0.945 (0.844, 1.000)	0.815 (0.711, 0.907)
	Pathological	0.916 (0.747, 0.996)	0.899 (0.712, 0.998)
Expert 1	Full dataset	0.964 (0.937, 0.983)	0.915 (0.887, 0.936)
	Normal	0.996 (0.976, 1.000)	0.913 (0.864, 0.939)
	TMJ Prosthetic	0.865 (0.776, 0.933)	0.936 (0.872, 0.977)
	Orthognathic	0.995 (0.970, 1.000)	0.866 (0.796, 0.917)
	Pathological	0.960 (0.881, 0.997)	0.969 (0.905, 1.000)
Expert 2	Full dataset	0.932 (0.899, 0.958)	0.874 (0.838, 0.904)
	Normal	0.950 (0.893, 0.984)	0.886 (0.828, 0.925)
	TMJ Prosthetic	0.856 (0.766, 0.927)	0.913 (0.841, 0.961)
	Orthognathic	0.947 (0.888, 0.982)	0.793 (0.704, 0.861)
	Pathological	0.959 (0.882, 0.997)	0.928 (0.842, 0.981)
Expert 3	Full dataset	0.872 (0.829, 0.907)	0.841 (0.801, 0.875)
	Normal	0.928 (0.864, 0.970)	0.887 (0.827, 0.925)
	TMJ Prosthetic	0.800 (0.697, 0.891)	0.885 (0.810, 0.934)
	Orthognathic	0.893 (0.821, 0.944)	0.786 (0.698, 0.856)
	Pathological	0.829 (0.709, 0.918)	0.798 (0.678, 0.885)

Table 2. Qualitative reproducibility results, in terms of the repeatability measure, for the Likert scoring of the radiologist's annotation and the DLS output. Results shown for each heterogeneity group and radiologist, as well as for the full dataset and average of the Experts. CI denotes the Bayesian credibility interval.

In terms of descriptive statistics, the two ratings of the three experts resulted in 1908 ratings, both for the radiologist and the DLS. 1893 (99.2%) canal paths from the radiologist and 1828 (95.8%) from the DLS were rated diagnostically fully usable. Taking the median rating of the Experts for each canal showed that 316 (99.4%) canals by the radiologist and 305 (95.9%) canals by the DLS were fully suitable for diagnosis, with 1 (0.3%) canal for the radiologist and 2 (0.6%) canals for the DLS uncertain, and with 1 (0.3%) canal for the radiologist and 11 (3.5%) canals for the DLS not fully diagnostically usable. The mean, standard deviation, and counts of the Likert ratings for each heterogeneity group are shown in Table 3.

The error types reported by the Experts were also aggregated by majority voting for each of the canals. The radiologist's annotations had errors for cases in TMJ Prosthetic group, specifically, 2 (2.9%) were short at mandibular foramen and 1 (1.4%) slightly off centre. The errors of the DLS were; fully missing canal for 2 (2.0%) cases from the Normal and 4 (4.3%) from the Orthognathic group, major parts missing in 1 (1.4%) case from TMJ Prosthetic group, short at mandibular foramen for 1 (1.5%) case from Pathological and 1 (1.4%) case from TMJ Prosthetic, and short at mental foramen for 2 (2.0%) from Normal and 2 (2.2%) from Orthognathic group.

In terms of the interobserver agreement of the Likert rating, the accuracy between Expert 1 and Expert 2 was found to be 94.81% and 95.91%, between Expert 1 and Expert 3 to be 91.98% and 91.51%, and between Expert 2 and Expert 3 to be 91.04% and 90.87%. The intraobserver agreement was found to be 99.37%, 95.91%, and 92.45% for Experts 1, 2 and 3, respectively.

Discussion

In this study, we have demonstrated the reproducibility capacity of a deep learning based automatic mandibular canal localisation system on a heterogeneous set of temporal clinical data. The DLS achieved both a high localisation performance and reproducibility in our systematic quantitative and clinical performance evaluation of multiple graders. In addition, the present work is based on an out-of-distribution dataset that highlights the evidence of the temporal generalisation capacity of the DLS.

In the quantitative analysis, the DLS produced good repeatability coefficient values for Normal, TMJ Prosthetic, and Pathological groups, while the Orthognathic group had in comparison worse mean, standard deviation, and repeatability coefficient value. This can be explained by the canals being affected by bilateral sagittal split osteotomy, which were generally difficult for the DLS. In this operation, the osteotomy lines go into the posterior mandibular canal areas, and after the ossification in the postoperative CBCT scans, the shapes, paths, and lengths of the mandibular canals may become modified. The postoperative mandibular canal can even have accessory openings to the surface of the mandible, and it can be asymmetrical to the contra-lateral mandibular canal. The generalisation performance for the BSSO cases is worse likely due to the lack of these cases in the training data. As for the overall generalisability results for SMCD, ASSD, and DSC, the DLS shows similar characteristics with the previous studies for the Normal, TMJ Prosthetic, and Pathological heterogeneity groups[13,16]. Specifically, the median SMCD values obtained for the DLS were lower than reported interobserver variability[16]. Additionally, the mean DSC was within the range of interobserver variability of 0.46–0.60 reported in the Supplementary section of that study. Furthermore, the DLS had large localisation and segmentation errors only on few canals that can be seen as outliers in the result figures. There is a similar proportion of outliers as in the previous study[16] and they are likely to be caused by some artefacts or difficult heterogeneity.

In the qualitative or clinical analysis, the radiologist had a higher probability for receiving the same Likert grading on two temporal segmentations, i.e. the repeatability measure, than the DLS on Normal and Orthognathic groups. On the Pathological group, the radiologist had a higher repeatability measure than the DLS with Expert 2, Expert 3, and on average, but lower repeatability with Expert 1. However, the DLS showed higher repeatability measure on the TMJ Prosthetic group for all the Experts and on average. Challenges in the TMJ Prosthetic group for the radiologist can be caused by metallic artefacts that affect the visibility of the canal (Fig. 2b). The patients of the group had been operated using an individually designed full TMJ-reconstruction, with condylar component that consists of titanium, and fossa component of Ultra-high molecular weight polyethylene, both fixed to the bone with titanium screws. These metallic structures cause major artefacts on the CBCT scans on the side with the operated mandible, degrading the visibility of the mandibular canal.

Heterogeneity	Marker	Mean	SD	N_0	N_1	N_2	N_3	N_4
Full dataset	Radiologist	3.94	0.27	0	3	12	72	1821
	DLS	3.84	0.65	35	9	35	75	1753
Normal	Radiologist	3.97	0.19	0	0	2	15	595
	DLS	3.86	0.62	12	0	8	21	571
TMJ Prosthetic	Radiologist	3.88	0.42	0	3	6	28	383
	DLS	3.92	0.38	0	4	2	18	396
Orthognathic	Radiologist	3.97	0.22	0	0	4	11	537
	DLS	3.71	0.90	23	3	21	17	487
Pathological	Radiologist	3.94	0.23	0	0	0	18	306
	DLS	3.90	0.39	0	2	4	19	299

Table 3. Mean, standard deviation (SD), and counts of all the Likert ratings (N_L for the Likert score L) grouped by the heterogeneity and the marker (the radiologist or the DLS).

Overall, the DLS showed slightly worse performance than the radiologist in terms of the average Likert score on the fully diagnostically usable canals. The DLS produced different types of errors, such as fully missing canals or major parts missing, while the radiologist's errors were mostly too short or slightly off centre canals. The error of too short canals can be expected, as the length of a mandibular canal is challenging to evaluate in the foramen mandibulae area. This is because of the anatomical features and subjective ending point of the canals[16]. The error of a canal being slightly off centre can be caused by an interpolation artefact or human error, when marking the canal manually[16]. There were small Likert score differences between the heterogeneity groups for both the radiologist's annotations and the DLS outputs. The group with the worst performance for the radiologist was TMJ Prosthetic group with the average Likert score of 3.88, while the worst performing group for the DLS was Orthognathic group with the average Likert score of 3.70. We found that there were no major differences between the majority of the qualitative results for the different heterogeneity groups. Indeed, most of the differences were caused by a few outlier cases. In terms of the patient level performance, the DLS turned out to have low variation between each scan of a patient, even for most of the cases that were imaged with different scanning devices and voxel spacings. Lastly, we evaluated the intra- and interobserver variability of the Likert scores given by the three experts in the two scoring sessions. We found out that there is a very strong overall intra- and interobserver reliability, but there are some differences between the Experts. Lower interobserver results can be explained by the subjective nature of the mandibular canal annotations, which is based on the radiologist's experience in the task[27].

There are some limitations in this study. First, we did not report all of the possible heterogeneities affecting the scans, such as difficult bone structure and movement of the patient. Second, the mandibular canal regions may also be affected by normal dental treatment during the follow-up period, which can cause changes to the nearby areas, such as new implants, dental crowns, endodontic materials, and orthodontic fixed appliances. These should be taken into consideration when designing future follow-up studies. In addition, further development is required to improve the robustness of the system for most anatomically, pathologically, or surgically difficult cases.

This work demonstrates the necessary capabilities of deep learning for longitudinal follow-up studies with different scanners or imaging parameters, and when there are anatomical or pathological changes in the patient's mandible. The next step in terms of a clinical validation will be the evaluation of the deep learning system under radiologist's supervision. Once validated and accepted for clinical use, the DLS can be used as a valuable tool in daily radiological, surgical, and dental practices, minimising potential complications in major and minor surgery. This reduces the time spent on computerised surgical planning in implantology and more complex maxillofacial surgery. In addition, it can be used to better visualise anatomical relationships for a variety of radiological challenges, including impacted teeth, odontogenic cysts and tumours, and malignant tumours, compared to static reconstructions. This visualisation would save time and effort in the preparation and reading of radiological reports.

Conclusion

The reproducibility of mandibular canal localisation with a deep learning system was evaluated on a heterogeneous dataset of temporal cone beam computed tomography scans. The reproducibility was found to be comparable to a radiologist in terms of quantitative and qualitative measures. In case of localisation and segmentation, the generalisation performance was found to be within or better than the interobserver variability suggested in the literature.

Data availability

The datasets used in model training, validation, and testing were provided by TAUH and CMU, and as such is not publicly available and restrictions apply to their use according to the Finnish law and General Data Protection Regulation (EU) and to the Thai law, respectively. The datasets used in this study is available from the corresponding author upon reasonable request and with permissions of TAUH and CMU.

Code availability

The code used for pre- and postprocessing and the deep learning techniques includes proprietary parts and cannot be released publicly. However, the approach can be replicated using the information from the cited literature and the evaluation measures with the equations in the "Methods" section and in Supplementary Information.

References

1. Agbaje, J. O. *et al.* Tracking of the inferior alveolar nerve: Its implication in surgical planning. *Clin. Oral Invest.* **21**, 2213–2220 (2017).
2. Bertl, K., Heimel, P., Reich, K. M., Schwarze, U. Y. & Ulm, C. A histomorphometric analysis of the nature of the mandibular canal in the anterior molar region. *Clin. Oral Invest.* **18**, 41–47 (2014).
3. Oliveira-Santos, C. *et al.* Visibility of the mandibular canal on CBCT cross-sectional images. *J. Appl. Oral Sci.* **19**, 240–243 (2011).
4. Rood, J. & Shehab, B. N. The radiological prediction of inferior alveolar nerve injury during third molar surgery. *Br. J. Oral Maxillofac. Surg.* **28**, 20–25 (1990).
5. Renard, F., Guedria, S., Palma, N. D. & Vuillerme, N. Variability and reproducibility in deep learning for medical image segmentation. *Sci. Rep.* **10**, 1–16 (2020).
6. Brown, J. *et al.* Basic training requirements for the use of dental CBCT by dentists: A position paper prepared by the European academy of dentomaxillofacial radiology. *Dentomaxillofacial Radiol.* **43**, 20130291 (2014).
7. Macleod, I. & Heath, N. Cone-beam computed tomography (CBCT) in dental practice. *Dent. Update* **35**, 590–598 (2008).
8. Liang, X. *et al.* A comparative evaluation of cone beam computed tomography (CBCT) and multi-slice CT (MSCT): Part I. on subjective image quality. *Eur. J. Radiol.* **75**, 265–269. https://doi.org/10.1016/j.ejrad.2009.03.042 (2010).

9. Recht, M. P. *et al.* Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur. Radiol.* **30**, 3576–3584 (2020).
10. Sunoqrot, M. R. *et al.* The reproducibility of deep learning-based segmentation of the prostate gland and zones on T2-weighted MR images. *Diagnostics* **11**, 1690 (2021).
11. Braun, M. J. *et al.* Dental and maxillofacial cone beam CT-high number of incidental findings and their impact on follow-up and therapy management. *Diagnostics* **12**, 1036 (2022).
12. Kim, H., Park, C. M. & Goo, J. M. Test-retest reproducibility of a deep learning-based automatic detection algorithm for the chest radiograph. *Eur. Radiol.* **30**, 2346–2355 (2020).
13. Jaskari, J. *et al.* Deep learning method for mandibular canal segmentation in dental cone beam computed tomography volumes. *Sci. Rep.* **10**, 1–8 (2020).
14. Cipriano, M. *et al.* Deep segmentation of the mandibular canal: a new 3D annotated dataset of CBCT volumes. *IEEE Access* **10**, 11500–11510 (2022).
15. Issa, J., Olszewski, R. & Dyszkiewicz-Konwińska, M. The effectiveness of semi-automated and fully automatic segmentation for inferior alveolar canal localization on CBCT scans: A systematic review. *Int. J. Environ. Res. Public Health* **19**, 560 (2022).
16. Järnstedt, J. *et al.* Comparison of deep learning segmentation and multigrader-annotated mandibular canals of multicenter CBCT scans. *Sci. Rep.* **12**, 18598. https://doi.org/10.1038/s41598-022-20605-w (2022).
17. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, (eds Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer International Publishing, Cham, 2015).
18. Lee, T.-C., Kashyap, R. L. & Chu, C.-N. Building skeleton models via 3-D medial surface axis thinning algorithms. *CVGIP: Graph. Model. Image Process.* **56**, 462–478 (1994).
19. Will, L. A. & West, R. A. Factors influencing the stability of the sagittal split osteotomy for mandibular advancement. *J. Oral Maxillofac. Surg.* **47**, 813–818 (1989).
20. Politis, C. *et al.* Visibility of mandibular canal on panoramic radiograph after bilateral sagittal split osteotomy (BSSO). *Surg. Radiol. Anat.* **35**, 233–240 (2013).
21. Barnhart, H. X. & Barboriak, D. P. Applications of the repeatability of quantitative imaging biomarkers: A review of statistical analysis of repeat data sets. *Trans. Oncol.* **2**, 231–235 (2009).
22. Nakagawa, S. & Schielzeth, H. Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biol. Rev.* **85**, 935–956 (2010).
23. Culp, S. L., Ryan, K. J., Chen, J. & Hamada, M. S. Analysis of repeatability and reproducibility studies with ordinal measurements. *Technometrics* **60**, 545–556. https://doi.org/10.1080/00401706.2018.1429317 (2018).
24. Muraki, E. A generalized partial credit model: Application of an EM algorithm. *ETS Res. Rep. Ser.* https://doi.org/10.1002/j.2333-8504.1992.tb01436.x *xd (1992).
25. Stan Development Team. Stan modeling language users guide and reference manual (2022). Version 2.30. https://mc-stan.org/
26. Riddell, A., Hartikainen, A. & Carter, M. pystan (3.0.0). howpublishedPyPI (2021).
27. Hamid, M. M. & Suliman, A. M. Diameter of the inferior alveolar canal - a comparative CT and macroscopic study of sudanese cadaveric mandibles. *J. Evol. Med. Dent. Sci.* **10**, 342–346 (2021).

Acknowledgements

Author contributions

J.J. conception, design of the work, data and annotation acquisition, interpretation of data, data annotation, wrote the main manuscript text; J.S. & J.Ja. conception, design of the work, ran the experiments, analysis, wrote the main manuscript text and produced figures; K.K. conception, interpretation of data, project management, corresponding author, wrote and polished the main manuscript text; H.M. interpretation of data and data annotation; A.H. project management, design of the work, analysis; O.S., V.V., V.M. project management and design of the work; S.P., S.N. project management, design of the work, data acquisition; All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-40516-8.

Correspondence and requests for materials should be addressed to J.J. or K.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.