Mathematical and Computational Applications

MDPI

Article

Linear and Non-Linear Regression Methods for the Prediction of Lower Facial Measurements from Upper Facial Measurements

Jacques Terblanche, Johan van der Merwe * and Ryno Laubscher

Department of Mechanical and Mechatronic Engineering, Stellenbosch University, Stellenbosch 7600, South Africa; 22548602@sun.ac.za (J.T.)
* Correspondence: jovdmerwe@sun.ac.za

Abstract: Accurate assessment and prediction of mandible shape are fundamental prerequisites for successful orthognathic surgery. Previous studies have predominantly used linear models to predict lower facial structures from facial landmarks or measurements; the prediction errors for this did not meet clinical tolerances. This paper compared non-linear models, namely a Multilayer Perceptron (MLP), a Mixture Density Network (MDN), and a Random Forest (RF) model, with a Linear Regression (LR) model in an attempt to improve prediction accuracy. The models were fitted to a dataset of measurements from 155 subjects. The test-set mean absolute errors (MAEs) for distance-based target features for the MLP, MDN, RF, and LR models were respectively 2.77 mm, 2.79 mm, 2.95 mm, and 2.91 mm. Similarly, the MAEs for angle-based features were 3.09°, 3.11°, 3.07°, and 3.12° for each model, respectively. All models had comparable performance, with neural network-based methods having marginally fewer errors outside of clinical specifications. Therefore, while non-linear methods have the potential to outperform linear models in the prediction of lower facial measurements from upper facial measurements, current results suggest that further refinement is necessary prior to clinical use.

Keywords: orthognathic surgery; mandible shape prediction; non-linear methods; 3D cephalometrics

Citation: Terblanche, J.; Merwe, J.v.d.; Laubscher, R. Linear and Non-Linear Regression Methods for the Prediction of Lower Facial Measurements from Upper Facial Measurements. *Math. Comput. Appl.* **2024**, *29*, 61. https://doi.org/10.3390/mca29040061

Academic Editors: Leonardo Trujillo and Oliver Schütze

Received: 14 June 2024
Revised: 20 July 2024
Accepted: 30 July 2024
Published: 31 July 2024

1. Introduction

Orthognathic surgery aims to address malocclusion or facial deformities in cases where the midface of a patient is anatomically correct [1]. A successful orthognathic treatment can notably improve a patient's quality of life and facial aesthetics [2,3]. This surgery typically does not entail the changing of morphology but rather of the relative positions between the maxilla, mandible and cranium [4]. It is the clinician's responsibility to establish this original dentofacial shape, estimate the final position, and determine the actions required to achieve this position [5]. These assessments are typically based on the clinician's experience, physical models, and established population normative measurements derived from cephalometric analysis. This can be time-intensive and surgeon-dependent [1,5,6]. As a complementary method, patient-specific methods provide a data-driven technique to guide the surgery using the patient's own geometry [1].

A traditional approach to evaluating the harmony of patient geometry involves the harmony box [7]. This method determines normal ranges for critical cephalometric angles, such as SNA, from a many-to-one relationship using multiple cephalometric angles. In this method, the standard errors derived from a linear regression model constructed for a specific population are used to establish a table of floating normative values for that population. A clinician can then compare a patient's measurements to the table and determine if orthognathic treatment is necessary and, if so, which specific treatment should be considered. However, limitations include a small selection of variables, no consideration of multi-dimensional target relations, and a dependence on the singular target feature selection. To address this, Bingmer et al. [8] extended the harmony box approach with a

multiple linear regression model. Here, they used a wider range of variables, multiple output variables, and data from a normal population. They demonstrated the viability of using the residual error values to determine the normality of unseen patient geometry. Using this model, if a patient's residual errors are large, it may indicate a non-harmonious relationship requiring orthognathic treatment.

In a broader application focused on the prediction of mandible measurements, Gillingham et al. [9] employed multiple linear regression models to predict missing mandible cephalometric measurements from remaining measurements that span the craniofacial complex. They compared their linear regression models against the mirroring of geometry across the symmetry plane and found that their linear regression model outperformed mirroring in 6 of 16 measurements. Cheng et al. [4] predicted reposition vectors based on cephalometric coordinates using linear ridge regression and a non-linear transformed-based neural network. They concluded that the neural network had better generalisation capability compared to their linear model. Wang et al. [1] developed a shape-based sparse coefficient regression model using cephalometric landmarks to predict mandibular and maxillary structure from midface geometry. These coefficients, based on the 3D locations of the landmarks, were then used to predict mandibular landmarks. Their model demonstrated good qualitative results with better normality than the true input data, suggesting more harmonious outputs. Similarly, Zhu et al. [10] also applied both landmark- and measurement-based approaches, using linear regression models to predict landmarks on the maxilla based on data from the mandible and vice versa. They noted that their results did not achieve sufficiently low error metrics for clinical application.

Based on the previous review of the most relevant literature [1,4,7–10], it seems that limited attention was given to the prediction of mandible morphology solely from that of the cranium or maxilla as described by Zhu et al. This may be beneficial for orthognathic applications where the lower facial shape may be inharmonious or for mandible reconstruction where the mandible may be absent or only partially present. Furthermore, not all studies focused on reproducing cephalometric measurements [1,4,10], which the authors consider necessary to describe mandible shape variation for comparison with the literature. In addition, limited attention was given to non-linear methods [4] in comparison with linear approaches [1,7–10]. Non-linear models may be able to better represent the poor linear relationship between lower and upper facial measurements [11], warranting further study. Finally, while not directly comparable, current predictions of the mandible from the morphology of the maxilla yield reconstruction accuracy ranging from approximately 2.9 to 7.9 mm [10], which may be insufficient for clinical applications assuming a target value of 2 mm [12].

Therefore, in this paper, regression models were constructed based on clinically relevant cephalometric measurements derived from CT scans of 155 healthy subjects. Specifically, this is the first study where multiple linear regression models were compared in terms of predicting cephalometric mandible measurements based on upper facial measurements against non-linear models. The non-linear models included a random forest, a multi-variate multi-layer perceptron, and a multi-variate mixture density network. Finally, the ability of the models to meet clinical thresholds of 2 mm and 2 degrees was evaluated.

The following is an overview of the current paper's layout. Section 2 discusses the materials and methods. It describes the collection and processing of CT scans into relevant cephalometric measurements, followed by model and hyper-parameter optimisation background theory and their implementation. Section 3 combines the results and their discussion. This section will initially present the hyper-parameter optimisation outputs and their insights. Subsequently, the optimised models are evaluated against each other. The results of the models are discussed in terms of adherence to clinical thresholds, an overall comparison of error metrics, the clinical interpretation of these results, and potential future work. The paper concludes with a summary of the findings and implications in Section 4.

2. Materials and Methods

2.1. Data Collection and Processing

A total of 627 DICOM (Digital Imaging and Communications in Medicine) format head and neck CT scans were obtained from the HNSCC (Head and Neck Squamous Cell Carcinoma) database [13]. This dataset is a publicly accessible repository, with all individuals anonymised and de-identified. In accordance with international standards and guidelines, ethical clearance for this project was granted by the Health Research Ethics Committee of Stellenbosch University, South Africa (Ethics Reference Number: X23/07/027). This database consists of typically older individuals diagnosed with HNSCC. Scans were included based on the following criteria: (1) an in-plane resolution of 0.488 mm and slice thicknesses of 1.25 mm; (2) the absence of obvious bone pathology in the maxillofacial region; (3) the presence of all necessary cephalometric landmarks; (4) minimal scan artefacts; and (5) the presence of contrast agents. Notably, to allow for a larger sample size, the current work did not exclude subjects based on malocclusion class, dentition, or sex. The final dataset's demographics are summarised in Table 1. A file listing the unique identifiers of the selected subjects is available in the Supplementary Materials.

Table 1. Demographics of included subjects.

Patients	Number of Subjects	Age (Years)
Male	133	58.98 ± 9.91
Female	22	56.73 ± 8.24
Combined	**155**	$\mathbf{58.66 \pm 9.70}$

All scans were processed using the open-source software 3D Slicer 5.2.2 [14] (https://www.slicer.org/). Scans were segmented by applying a density threshold to isolate osseous structures, followed by manual removal of scanning artefacts such as streaking. Targeted and global Gaussian smoothing filters were then applied to reduce the voxelated appearance and remove minor discontinuities before exporting the resultant masks to face-vertex meshes in STL format. Following segmentation, the STL files were imported into MATLAB, where a previously developed in-house script was used to place 3D landmarks on the mandible, maxilla, and cranium. Landmark definitions were based on those of Bayome et al. and Gillingham et al. [11,15] and are listed in Tables A1 and A2 in the Appendix A. Figure 1 illustrates the placement of the mandible landmarks, and Figure 2 similarly shows the placement of the cranium and maxilla landmarks.

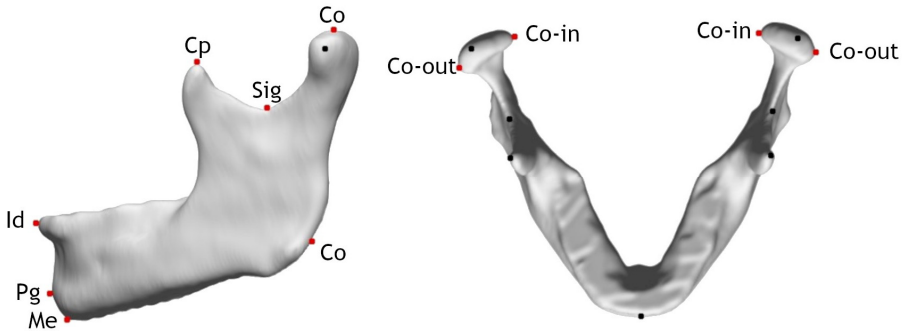


Figure 1. Mandible landmarks.

Cephalometric measurements were derived based on the coordinates of the obtained landmarks, again based on work by Bayome et al. and Gillingham et al. [9,15]. Additional upper facial measurements incorporating the Porion landmark were also defined. All measurements were extracted from the defined 3D landmark coordinates using a Python script. The mandible measurements are listed in Table 2, and the cranium and maxilla measurements are listed in Table 3.

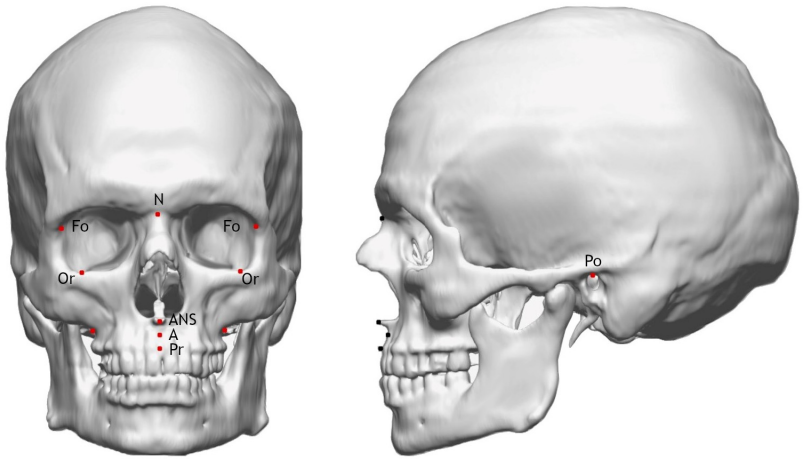


Figure 2. Cranium and maxilla landmarks.

Table 2. Definitions of mandible measurements.

Mandible	
Measurement	**Description**
Go-Sagittal	Distance from Gonion to sagittal plane
Co-Sagittal	Distance from Condylion to sagittal plane
Co-Sig	Distance from Condylion to Sigmoid notch
ID-Me	Distance from Infradentale to Menton
Go-Me	Distance from Gonion to Menton
Co-in–Co-out	Distance from the innermost condylar point to outermost condylar point (also known as condylar width)
Me angle	Angle spanned by the right and left Gnathion-Me
Me-Go-Co	Angle spanned by Menton, Gonion and Condylion.
Co-Go-FH	Angle spanned by Condylion, Gonion and the Frankfurt Horizontal plane

Table 3. Definitions of cranium and maxilla measurements.

Cranium and Maxillary	
Measurement	**Description**
N-ANS	Distance from Nasion to Anterior Nasal Spine
FO-FO	Right-to-left distance between Frontomalare Orbitales
ZA-ZA	Right-to-left distance between Zygomaxillare Anteriores
Or-Or	Right-to-left distance between Orbitales
SNA	Angle spanned by Sella, Nasion and A-point
N-Po-Pr Angle	Angle spanned by Nasion, Porion and Prosthion
Fo-Or-M	Angle spanned by the Frontomalare orbitale, Orbitale and Maxillary point
Po-Pr	Distance between Porion and Prosthion
Po-N	Distance between Porion and Nasion

To predict lower facial measurements from upper facial measurements, all available upper facial measurements were used as input features. For the ideal lower facial measurements, facial symmetry was assumed by disregarding the right-sided measurements and keeping only the left-sided measurements. This resulted in a total of 13 input features (upper facial measurements), with 9 target output features (lower facial measurements) listed, respectively, in Tables 2 and 3. To assess landmark placement repeatability, 10 scans were randomly selected for an intra-class correlation (ICC) test with two raters. Each individual, known as a rater, annotated 10 scans (i.e., the subjects) based on the defined landmarks. The consistency was assessed for multiple raters with assumed two-way mixed effects. Accordingly, Equation (1) [16] was used to determine the score, ICC_{score}, for each measurement. In this equation, *MSBS* represents the mean squared errors between subjects,

and *MSE* represents the total mean squared error. The reader is referred to the supporting information provided by Liljequist et al. [17] for further implementation details.

$$ICC_{score} = \frac{MSBS - MSE}{MSBS} \tag{1}$$

The data were randomly split into training, validation, and test datasets with a respective 70-20-10 split. The training set was composed of 108 individuals. The validation set, used for hyper-parameter optimisation and model selection, included 31 individuals. Finally, the test set for the final model comparison contained the remaining 16 individuals. To reduce the impact of the data's magnitude ranges and different unit types, the training set was scaled with a minmax scaler, as shown in Equation (2). Here x_{ij} represents the ith subject from the jth measurement type. Similarly, x^*_{ij} refers to the same data point but scaled with the min-max function. Finally, $\boldsymbol{x}_j$ represents all values of the jth measurement. The training data's scaling factor was also applied to all instances in the validation and test sets.

$$x^*_{ij} = \frac{x_{ij} - min(\boldsymbol{x}_j)_{train}}{max(\boldsymbol{x}_j)_{train} - min(\boldsymbol{x}_j)_{train}} \tag{2}$$

2.2. Model Construction

All models are implemented in the Python programming language. The Ordinary Least Squares Regression (OLS) and Random Forest (RF) models were each implemented as a model for each output measurement using the scikit-learn package [18]. The Multi-Layer Perceptron (MLP) and Mixture Density Network (MDN) were implemented with multiple target outputs using the PyTorch package [19]. The neural networks were constructed to favour training logic simplicity and to determine whether standard implementations are sufficient. Therefore, constant learning rates were used, and no regularisation techniques were implemented. The models were trained for 4000 epochs. The hyper-parameters were optimised using Bayesian optimisation (BO) via the BoTorch implementation from the Ax platform [20,21]. Refer to Section 3.1 for details regarding specific hyper-parameters. The following is a brief overview of the construction and general theory behind the methods used.

2.2.1. Ordinary Least Squares Regression

Ordinary least squares (OLS) regression is a standard technique to determine the linear relationship between a set of input and output features. This technique is commonly applied in estimating mandibular morphology [4,7–10]. Moreover, the OLS model, known for its high bias (linear assumption) but low variance, may underfit complex relationships. However, it also ensures consistent performance across datasets and is robust against overfitting [22]. Therefore, given its widespread use in estimating mandibular morphology and its functionality as a simple regression model, the OLS model will serve as a baseline to compare against more complex models.

In this project, an OLS model was constructed for each output measurement; therefore, the number of models constructed was equal to the number of output features, $n_{outputs}$. Each OLS model adhered to the following described structure but predicted a different single output feature.

The OLS estimator, shown in Equation (3), is a set of coefficients that minimizes the sum of squared vertical distances between the fitted line and the data points [23]. Note that the equations are adapted for matrix notation. Here, $\mathbf{X}$ is the input training data matrix of the general size $n_{subjects} \times n_{inputs}$, $\mathbf{X}^T$ the transposed input data matrix, and $\boldsymbol{y}$ is a vector containing the true training dataset target features of size $n_{subjects}$. For clarity, $n_{subjects}$ refers to the number of subjects in the dataset (training, validation, or testing) and n_{inputs} refers to the number of input features. This convention is followed throughout this paper.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y} \tag{3}$$

The resultant OLS estimator, $\hat{\boldsymbol{\beta}}$ with size n_{inputs}, can be used along with the test input data, $\mathbf{X_{test}}$, to predict previously unseen target values from the testing dataset, denoted $\hat{\boldsymbol{y}}$, as in Equation (4).

$$\hat{\boldsymbol{y}} = \mathbf{X_{test}}\hat{\boldsymbol{\beta}} \tag{4}$$

The model inherently assumes a linear relationship, as it predicts the target feature through a linear combination of input features. However, it is possible to extend this technique so that it can capture some non-linear relationships. This can be achieved by applying a dimensional transform to the input data. The linear model is thereafter applied to this transformed space. One such method is the z-space feature transform, as represented in Equation (5) [24]. This transform involves constructing a new dataset, $\mathbf{X}'$, by adding columns where each column is $\mathbf{X}$ raised to successive powers from 1 up to k. The transformed dataset, now of size $n_{subjects} \times \left(k \times n_{\,outputs} + 1\right)$ is then similarly used to determine a transformed coefficient, $\hat{\boldsymbol{\beta}}'$. This is subsequently applied in Equation (4) to predict new data.

$$\boldsymbol{X}' = \left(1, \boldsymbol{X}^1, \ldots, \boldsymbol{X}^k\right) \tag{5}$$

2.2.2. Multi-Layer Perceptron

The multi-layer perceptron (MLP) is a relatively non-complex neural network architecture. If implemented with non-linear activation functions, it can approximate any continuous function [25], giving it low bias but high variance. Generally, this also makes it prone to overfitting. However, these issues can be mitigated by hyper-parameter optimisation (implicit regularisation) and various explicit regularisation techniques [22]. Despite its popularity across machine learning, the MLP model has seen less frequent usage in estimating mandible cephalometric measurements. Therefore, the MLP is included as a non-linear, representative basic neural network architecture for its potential in bias-variance control and for its relative absence in cephalometric measurement prediction. To this end, a single MLP was implemented to learn from multiple input measurements to predict multiple output measurements.

The MLP generates model outputs based on the provided weights and the input data through a forward-pass loop. For a general case, this forward-pass propagation is shown in Equation (6) [23]. Here the input measurements are represented by $\boldsymbol{x}$ with a size $n_{batch} \times n_{inputs}$ and the estimated output is represented by $\hat{\boldsymbol{y}}$ with a size $n_{batch} \times n_{outputs}$. This provided forward-pass equation is adapted to matrix notation and simplified with biases omitted. In addition, the current implementation assumes constant neuron counts, $n_{neurons}$, per layer and a batch input with size n_{batch}.

$$\hat{\boldsymbol{y}}(\boldsymbol{x}, \mathbf{W}) = \Phi_n\left(\mathbf{W}^{(n)}\Phi_{n-1}\left(\mathbf{W}^{(n-1)} \ldots \Phi_1\left(\mathbf{W}^{(1)}\boldsymbol{x}\right)\ldots\right)\right) \tag{6}$$

The network consists of an input layer with weights matrix $\mathbf{W}^{(1)}$ sized $n_{inputs} \times n_{neurons}$. This layer connects to m hidden layers, each characterised by an activation function, Φ, and a weights matrix, $\mathbf{W}$, sized $n_{neurons} \times n_{neurons}$. For regression tasks, the output layer utilizes a linear activation function, Φ_m, and a weights matrix, $\mathbf{W}$, sized $n_{weights} \times n_{outputs}$. The activation functions between the hidden layers generally involves non-linear transformations. These functions enable the neural network to learn non-linear relationships. A visual representation of an MLP implementation is provided in Figure 3.

In training the model, the model evaluates its output using a loss function. In this implementation, the loss function is the root mean squared error, as defined by Equation (7). In this equation, $\boldsymbol{y}$ again represents the true values as per the training or validation dataset.

$$\mathcal{L}_{MSE} = \frac{1}{n_{batch}}||\boldsymbol{y} - \hat{\boldsymbol{y}}||_2^2 \tag{7}$$

Using the backpropagation algorithm [26], the loss output is propagated through the neural network to compute the gradients of the loss with respect to the weights, **W**. The gradient information along with an optimiser, is used to update the weights of the network. In the present study, Adaptive Moment Estimation (Adam) [27] was chosen as the optimiser. This optimisation process then iterates over the provided dataset in batches to minimise the loss function. Each epoch is regarded as a complete pass of the dataset. Within the epochs, the data are divided into batches. On each pass of a batch through the network, the weights are updated based on the calculated gradients.

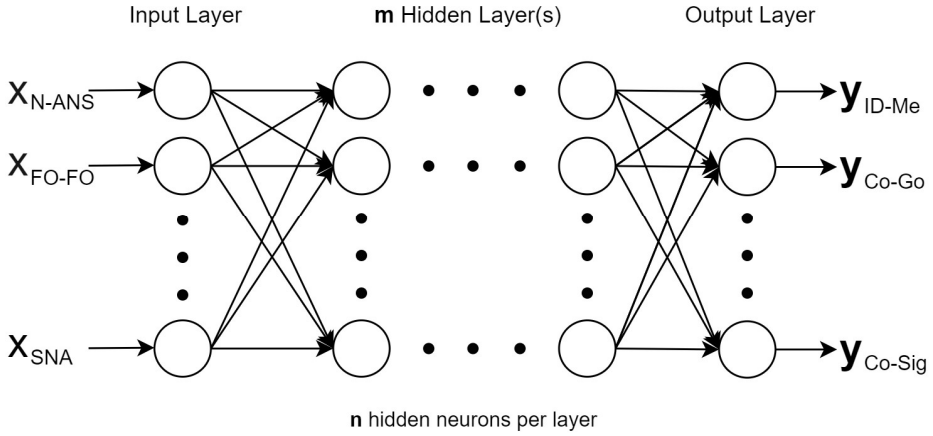


Figure 3. Neural network layout for a MLP with m hidden layers, each with n neurons per hidden layer. Not depicted on the diagram are the activation layers, neuron weights, and biases between neuron connections.

2.2.3. Mixture Density Network

The mixture density network (MDN) [28] has a similar architecture to the MLP; however, the mixture density network extends the MLP by predicting multiple ($n_{mixtures}$) distributions per target feature rather than only the expected value. This ability allows the MDN to model multi-modal behaviour [23], which may exist in skull morphology. Furthermore, it uniquely provides a measure of uncertainty (standard deviation) when applied to unseen data. Finally, to the authors' knowledge, no previous study has implemented a MDN for the prediction of cephalometric measurements. Therefore, in this paper, the MDN serves as a representation of a multi-modal extension of a neural network-based model and as the first implementation of a MDN in estimating cephalometric measurements.

The MDNs feed-forward propagation equation is shown in Equation (8), where the output $\boldsymbol{\theta}$, comprises of means ($\boldsymbol{\mu}$) with size $n_{batch} \times n_{outputs} \times n_{mixtures}$, standard deviations ($\boldsymbol{\sigma}$) with size $n_{batch} \times n_{outputs} \times n_{mixtures}$, and mixing coefficients ($\boldsymbol{\pi}$) with size $n_{batch} \times n_{mixtures}$. The network layout is also visually illustrated in Figure 4.

$$\boldsymbol{\theta}(\boldsymbol{x}, \mathbf{W}) = \Phi_{\mathrm{n}}\left(\mathbf{W}^{(n)}\Phi_{\mathrm{n}-1}\left(\mathbf{W}^{(n-1)} \ldots \Phi_1\left(\boldsymbol{W}^{(1)}\boldsymbol{x}\right) \ldots\right)\right) \tag{8}$$

As outlined in Bishop [28], to implement this model, constraints are required on all mixing coefficients and standard deviations. The mixing coefficients should sum to 1. This can be accomplished by applying the soft-max equation to each mixing coefficient π_k representing the weighting factor for the kth distribution, as shown in Equation (9). This function converts a set of unnormalised scores into a probability distribution. The second constraint is that the standard deviations are strictly positive, which can be enforced by applying an exponential function such as Equation (10). This transforms any real-valued inputs into positive values. Finally, the estimated mean value can be taken as in Equation (11).

$$\pi_k = \frac{\exp(\theta_{\mathbf{k}}^{\boldsymbol{\pi}})}{\sum_{k=1}^{K}\exp(\theta_{\mathbf{k}}^{\boldsymbol{\pi}})} \tag{9}$$

$$\sigma_k = \exp(\theta_k^{\sigma}) \tag{10}$$

$$\mu_k = \theta_k^{\mu} \tag{11}$$

In combining these parameters, a probability density function can be formed for K independent and isotropic Gaussian distributions conditioned with means, $\boldsymbol{\mu}$ and isotropic variances $\boldsymbol{\sigma}^2$. By using isotropic variances, the model assumes that the uncertainty is uniform across all output measurements for each component of the mixture. The related probability density function, shown in Equation (12), can be incorporated into the model's loss function. The negative loss-likelihood function, Equation (13), is formed by taking the negative log of the probability density function.

$$P(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}, \mathbf{W}) \mathcal{N}_k\Big(\boldsymbol{y} \,\Big|\, \mu_k(\boldsymbol{x}, \mathbf{W}), \sigma_k^2(\boldsymbol{x}, \mathbf{W})\mathbf{I}\Big), \tag{12}$$

$$E(\mathbf{W}) = -\sum_{n=1}^{n_{batch}} \log \sum_{k=1}^{K} \pi_k(\boldsymbol{x_n}, \mathbf{W}) \mathcal{N}_k\Big(\boldsymbol{y_n} \,\Big|\, \mu_k(\boldsymbol{x_n}, \mathbf{W}), \sigma_k^2(\boldsymbol{x_n}, \mathbf{W})\mathbf{I}\Big) \tag{13}$$

In this equation, $\boldsymbol{y_n}$ with size $n_{outputs}$, represents our target true measurements for the nth subject in the batch. Similarly, $\boldsymbol{x_n}$ with size n_{inputs}, refers to the input measurements for the nth subject. Furthermore, $\mathbf{I}$ is the identity matrix. This negative loss likelihood function can then be minimised in the training loop.

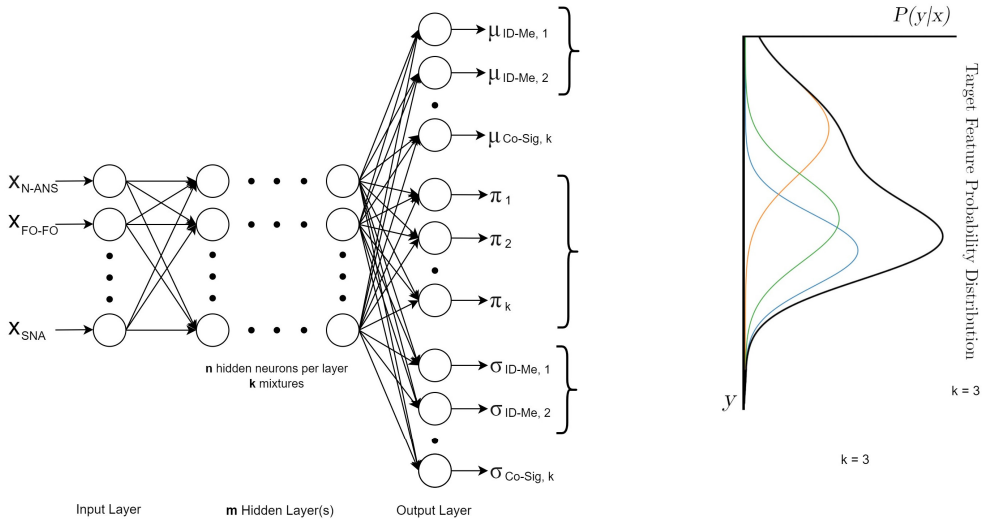


Figure 4. Neural Network layout for a MDN with $n_{mixtures}$ (k) distributions, m hidden layers, each with $n_{neurons}$ neurons per hidden layer. Not depicted on the diagram are the activation layers, neuron weights and biases between neuron connections. The plot on the right depicts an example of the predicted weighted distribution (bolded line) for a target feature with $n_{mixtures}$ = 3. The lines with colour represent all individual predicted distributions for that feature.

Similar to the Multilayer Perceptron (MLP), the loss function and the backpropagation algorithm are employed to propagate the loss backward through the network. The weights were again optimized using the Adam gradient-descent-based optimiser [27]. This updates the network weights, **W**, after each batch pass.

2.2.4. Random Forest

The random forest (RF) model for regression comprises an ensemble of multiple regression decision trees. RF models provide the capacity to learn non-linear relationships while also being computationally inexpensive and having higher interpretability in the form of feature importance [29]. Importantly, tree-based models have different inductive

biases compared to neural-network-based models. Therefore, the RF model is included as a representative tree-based and inexpensive non-linear model.

In this implementation, a RF model with multiple inputs is constructed for each output measurement. The decision tree implementation from Sci-Kit Learn is based on the classification and regression trees (CART) algorithm [30]. In CART-based regression trees, the dataset is recursively split by solving a minimisation problem aimed at reducing the overall loss function, denoted as G. This is represented as follows:

$$\theta^* = \text{argmin}_\theta G(\mathbf{Q}_m, \theta) \tag{14}$$

Here $\mathbf{Q}_m$ represents a subset of the data at node m, consisting of the input measurements $\mathbf{X}$ of size $n_{subjects} \times n_{inputs}$ and their respective target output measurements, $\boldsymbol{y}$ with size $n_{subjects}$. The parameters are represented with θ (of size $n_{paramaters}$). The data subsets, along with the parameters govern the splitting criterion, that minimises the total loss, G. This minimisation leads to data entries with similar target values being grouped together.

In model training, the CART implementation employs binary splits for each leaf, dividing data entries into a left dataset, $\mathbf{Q}_{\text{m}}^{\text{l}}(\theta)$, and a right dataset, $\mathbf{Q}_{\text{m}}^{\text{r}}(\theta)$, based on the selected split value. For the input features, the algorithm evaluates all possible splits, m, to identify those that maximise the reduction in loss. The loss is calculated between the target output and the split's mean output value for that measurement. The calculation of the total loss is the weighted sum of the losses of each partition, as shown in Equation (15).

$$G(\mathbf{Q}_m, \theta) = \frac{n_m^l}{n_m} H(\mathbf{Q}_m^l(\theta)) + \frac{n_m^r}{n_m} H(\mathbf{Q}_m^r(\theta)) \tag{15}$$

In this equation, the variables n_m^l and n_m^r represent the number of data entries in the left and right datasets, $\mathbf{Q}_{\text{m}}^{\text{l}}(\theta)$ and $\mathbf{Q}_{\text{m}}^{\text{r}}(\theta)$, as determined by the splitting of $\mathbf{X}$ and $\boldsymbol{y}$ at node m. This RF implementation used the mean squared error for its loss calculation. The related loss calculation for each data subset is shown in Equation (16).

$$H(\mathbf{Q_m}) = \frac{1}{n_m} ||\boldsymbol{y} - \hat{y}_m||_2^2 \tag{16}$$

In this equation, $\hat{y}_m$ represents the mean of the subjects for the specified upper facial measurement in the selected subset. This equation quantifies the error as the mean squared deviation of the actual targets from the subset's mean. The aforementioned steps are recursively applied to all training data until a specified stopping criterion is met. Once the tree is fully developed, each node (or split) is conditioned based on the training data, which segments the data into leaf end nodes. Each leaf node is associated with a $\hat{y}_m$ value.

In predicting unseen data, the tree determines the appropriate leaf node using the new input measurements and the previously conditioned nodes. The estimated measurement output value, $\hat{y}$, is assigned the $\hat{y}_m$ value associated with the determined leaf node.

For the current work, tree depth was selected as the stopping criterion. The resultant tree then forms part of the specified number of estimators in the random forest. A random forest regressor takes the average output of all its trees as the forest's output. Finally, Sci-Kit Learn's implementation also uses the concept of bagging. This is where each regression tree in the network is trained on a unique set of bootstrapped samples of the training dataset. For more implementation details, see the Sci-kit Learn [18] documentation.

2.3. Model Hyper-Parameter Optimisation

2.3.1. Grid-Search

A two-step approach for hyper-parameter optimisation was implemented. First, a grid search was used to provide insight regarding parameter effects and to subsequently decrease the search space. A grid search is an exhaustive search where models are trained for a cartesian product of sets of hyper-parameters. However, this approach is computationally

expensive for high-dimensional search spaces and has a fixed resolution [31]. The latter is not ideal for the optimisation of continuous parameters such as a neural network's learning rate. Bergstra and Bengio [31] suggested that adaptive learning models and random searches can more effectively find optimal model hyper-parameters.

2.3.2. Bayesian Optimisation

For the second step, Bayesian optimisation (BO) was implemented for adaptive hyper-parameter optimisation. The Bayesian search was applied to a reduced parameter space, as determined by the grid search, to find the optimal model parameters. For a detailed review of BO, refer to Shahriari [32]. Here, a high-level overview of BO is provided, as implemented in Ax-platform [21] using the BoTorch model [20]. The Ax-platform's mixed BO model integrates Sobol sequence sampling [33] to generate initial values, which are then used to establish a prior for BO.

BO is an optimisation algorithm with two main components, namely a surrogate and an acquisition function. The surrogate function is a less computationally expensive approximation of the objective function. This allows estimates of model performance for different hyper-parameter sets without direct evaluation of the objective function. One common implementation of the surrogate function is Gaussian Process Regression (GPR) [32]. GPR enables the model to estimate both the output value of a function and the uncertainty associated with that estimate. This uncertainty prediction is then used in the acquisition function.

The acquisition function determines which areas of the hyper-parameter space are expected to improve model performance. It does this by evaluating regions in the hyper-parameter space with high uncertainty or potential information utility. In the current work, the objective function would be the loss function. For this case, the expected improvement (EI) function as implemented in Ax-platform is shown in Equation (17).

$$\mathrm{EI}(\boldsymbol{x}) = \mathbb{E}[\max(f^* - f(\boldsymbol{x}), 0)] \tag{17}$$

Here f is the surrogate function and f^* is the smallest loss observed. The EI function determines the expected value of hyper-parameters, providing the largest decrease in loss. The parameters with the highest expected improvement are then used in the computationally expensive objective function. The objective function then computes the true performance of these hyper-parameter sets. Subsequently, the results are used to update the GPR model. This process is repeated for a specified number of trials, allowing the model to dynamically adapt and converge to a solution.

3. Results and Discussion

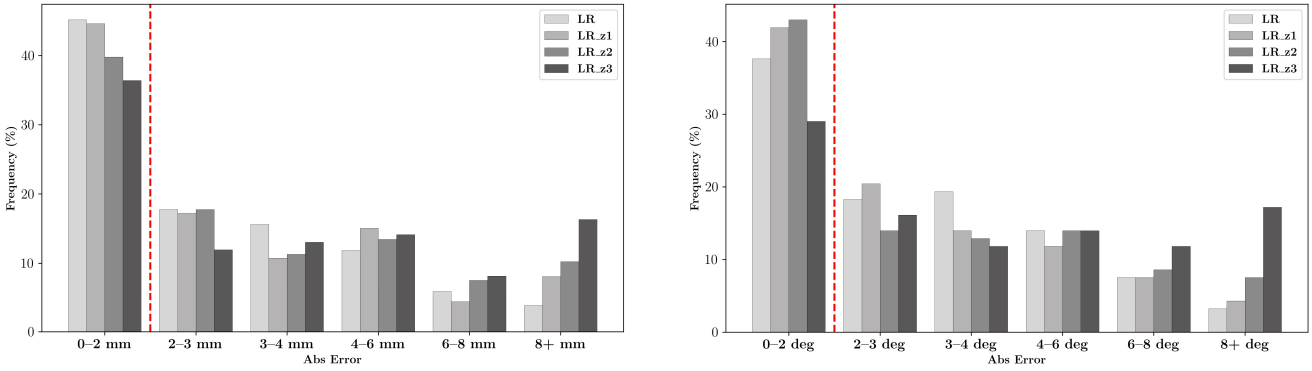
3.1. Model Choices and Hyper-Parameter Optimisation

All model selection choices in this section were based on the validation dataset. The base OLS model was compared against feature-transformed OLS models. All other models underwent hyper-parameter optimisation. Initially, a grid search was applied to the MLP. The resultant grid search outputs were then used to reduce the hyper-parameter optimisation space. With this reduced hyper-parameter space, the MLP underwent Bayesian optimisation. The optimised MLPs parameters were transferred to the MDN. This was followed by a minor grid search on the MDN to determine a viable range of mixture coefficients. Finally, Bayesian optimisation was applied to the MDN with the reduced hyper-parameter space. Due to lower computational costs, the random forest was directly optimised with Bayesian optimisation with a large hyper-parameter space.

3.1.1. Linear Regression

The standard OLS linear regression (LR) was compared to OLS with z-space transforms. The transforms applied include k = 2 (LR_z1), k = 3 (LR_z2), and k = 4 (LR_z3). Results for distance and angle measurements are reported separately. For the distance-based features, the mean absolute error (MAE) for each model based on the validation set was

2.78 mm, 3.10 mm, 3.59 mm, and 5.83 mm for the LR, LR_z1, LR_z2, and LR_z3, respectively. Similarly, for angle-based features, the validation MAE values were 3.00°, 3.03°, 3.39° and 5.93°, respectively. These results indicate that the base LR model, without transforms, had the lowest error rates. The differences in errors were more pronounced for distances than for angles. Given the clinical importance of maintaining errors within 2 mm or 2° within applications for virtual surgery preparation [12], the results were also assessed with error frequency diagrams. The frequency of prediction errors for the base OLS and for the transformed models is plotted for distances and angles, respectively, in Figures 5a and 5b.



(**a**) (**b**)

Figure 5. Frequency of absolute errors for LR models based on predicted measurements from validation dataset. The dashed line indicates the clinical threshold of 2 mm/°: (**a**) Absolute errors for distances. (**b**) Absolute errors for angles.

Figure 5a indicates that the base LR and the closely following LR_z1 model had the highest number of distance errors adhering to clinical requirements. The plot also demonstrates that the frequency of absolute distance errors decreased as higher-dimensional transforms were applied. Conversely, Figure 5b reveals that this trend did not extend to angular measurements. Here, LR_z1 and LR_z2 had the lowest number of errors below 2 degrees, whereas the higher k = 4 (LR_z3) transform still performed the worst. This may indicate that the relationship between input and the target angle features is more complex than that of the distances. Additionally, across both graphs, the transformed models showed a notably higher number of errors greater than 8 mm/°. These extreme values might explain the discrepancy between model performance regarding MAE and the frequency of errors under 2 mm/°. It is possible that this is because of overfitting. As the data increases in dimension, the models start overfitting on the training data [34], which leads to worse validation performance. Given these results, the base LR model, which demonstrated the lowest MAE values, a comparable number of errors below 2 mm/°, and the fewest extreme values were selected for further comparison.

3.1.2. Multi-Layer Perceptron

The MLP underwent a grid search to determine the effect of changing the hyper-parameters and to allow for a more focused Bayesian search. The grid search hyper-parameters are shown in Table 4.

Table 4. Hyper-parameter search space for multi-layer perceptron grid-search.

Hyper-Parameter	Values
Neuron count per layer	8, 32, 64, 128
Number of hidden layers	1, 2
Activation function	ELU, ReLU, LeakyReLU
Learning rate	Log range from 1×10^{-7} to 1×10^{-2}
Batch Size	8, 16, 32

To start analysing the grid search, the LeakyReLU activation function was chosen, and the batch size was set to 16. With these parameters, the learning rates for each neural network configuration are plotted against the final validation loss in Figure 6. The results indicate that the neural network configurations had the lowest average validation loss when the learning rate was 1×10^{-5}. Therefore, this value was selected as the model's learning rate for further analysis.

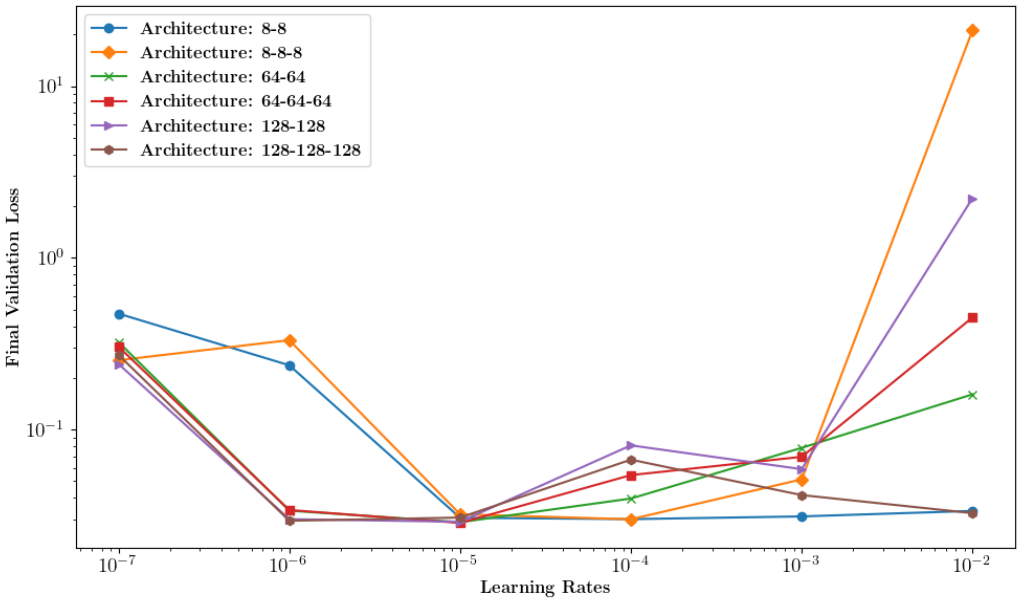


Figure 6. MLP grid search: Learning rates against validation loss for different neural network architectures (Act Function: LeakyReLU, LR: 1×10^{-5}).

Now, the neural network configurations for different batch sizes are plotted against the final validation loss in Figure 7. This plot shows that all models had similar performance across batch sizes, with the batch size of 8 generally having slightly better performance. As a result, for subsequent steps, the batch size was kept at 8.

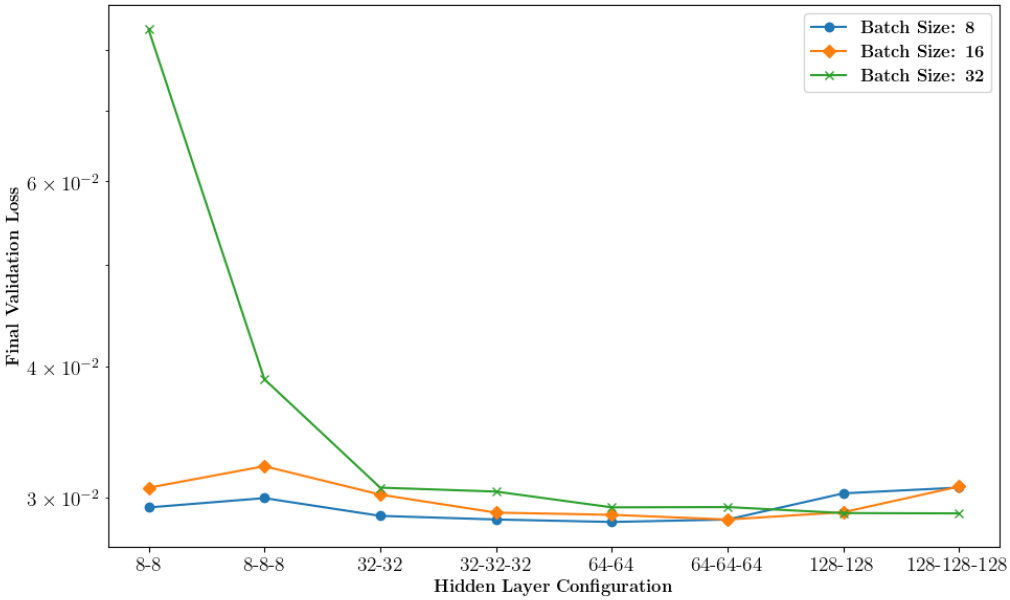


Figure 7. MLP grid search: Neural network architectures against validation loss for different batch sizes (Act. Function: LeakyReLU, Batch Size: 16).

In the final stage of the grid search, using the previously determined parameters, Figure 8 displays the validation loss for various activation functions plotted against different architectural configurations. This figure demonstrates that the ELU activation function had the overall most stable performance across all configurations. The ELU function also had the lowest instance of validation loss at a 32-32-32 configuration. This may be explained by ELU having the ability to centre the activations, similar to batch normalisation, and ELU being more robust to noise [35].

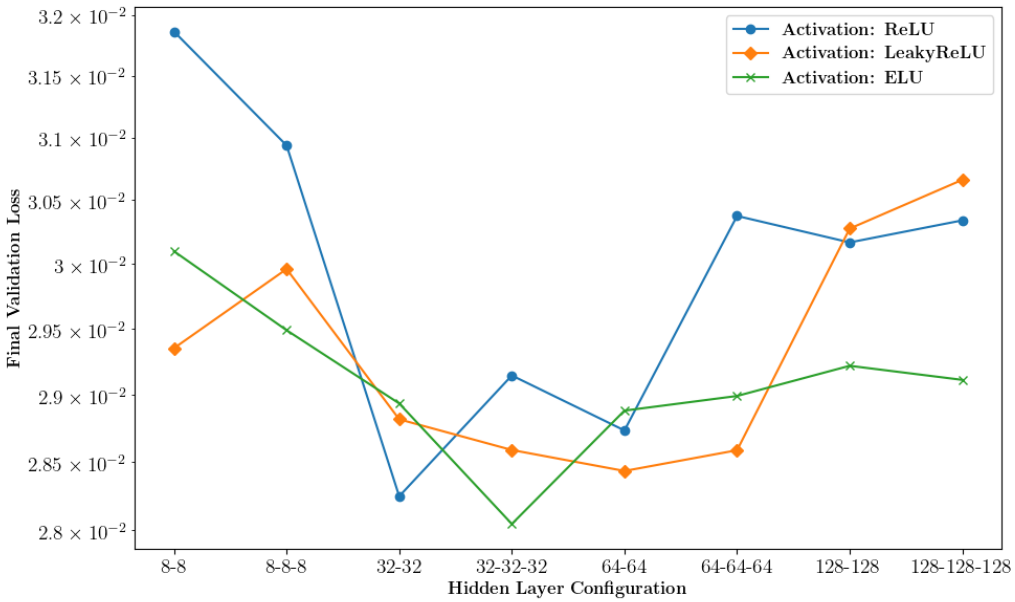


Figure 8. MLP grid search: Neural network architectures against validation loss for different activation functions (Batch Size: 16, LR: 1×10^{-5}).

Furthermore, the data suggest that increasing the neuron count beyond 64 did not improve performance. Additionally, the plot did not reveal any notable performance differences between architectures with two or three layers. It has been suggested that neural network depth is associated with its ability to learn more complex relationships [36]. Therefore, if non-linearity in the data were more significant relative to the noise, then it is speculated that there would be a trend with lower final validation loss for 3 layers compared to 2 layers.

Based on the grid search results, the parameters listed in Table 5 are the reduced hyperparameter search space for Bayesian optimisation (BO). With this search space, the Sobol-BO algorithm was applied for 200 trials. Of those 200, the Ax-platform determined that 6 Sobol sequencing trials were required to initially specify a prior. The parameter set with the lowest validation loss, trial 89, is also included in Table 5. In Figure 9, the BO trials are plotted against the final validation losses. Further hyperparameter tuning, with the current hyper-parameter space, did not appear to be computationally practical, as increasing the trial count did not lead to a noticeable reduction in loss. It is also apparent that the optimisation algorithm quickly converged to a minimal value and that various parameter combinations produced similar loss outcomes. It is speculated that these minima are relatively shallow and broad.

Table 5. Hyper-parameter search space for MLP Bayesian-search.

Hyper-Parameter	**Values**
Neuron count per layer	8, 32, 64
Number of hidden layers	2, 3
Activation function	ELU
Learning rate	Log range from 1×10^{-6} to 1×10^{-3}
Batch Size	8, 16
Determined Hyper-Parameters	**Values**
Neuron count per layer	32
Number of hidden layers	2
Activation function	ELU
Learning rate	6.62×10^{-6}
Batch Size	8

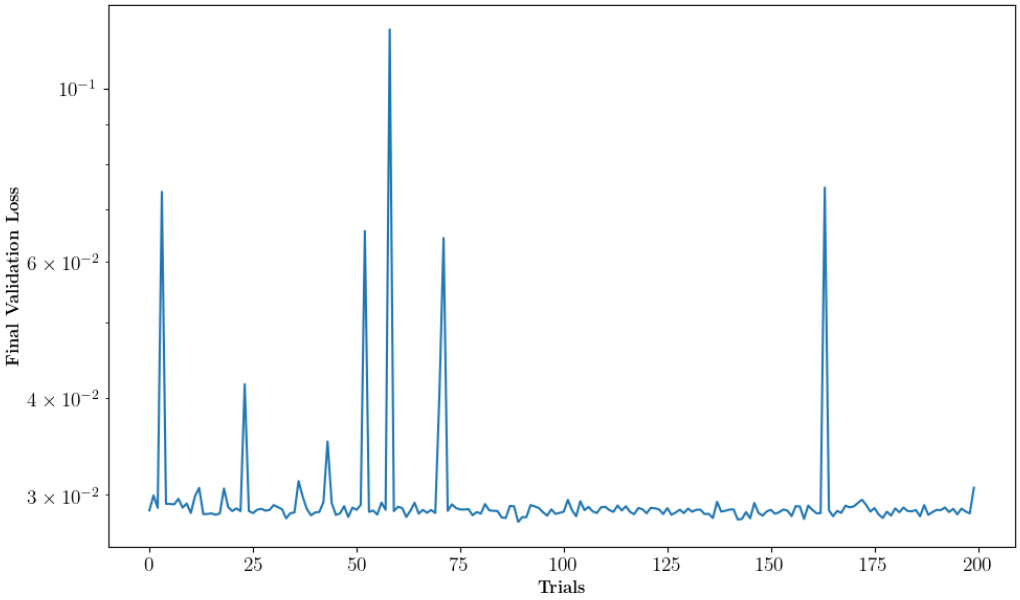


Figure 9. MLP BO: Bayesian Optimisation trials against final validation loss.

3.1.3. Mixture Density Network

The final grid-search parameters from the MLP were transferred to the MDN. With these parameters, a small grid search was then completed to optimise the mixture count and batch size. The results of this grid search are shown in Figure 10. The results suggest that the number of mixtures did not have a relationship that was easily distinguishable from noise. However, there was a significant impact in changing the batch size, which prompted a full BO search.

The hyper-parameter search space and final outputs of the Bayesian search (trial 184) are shown in Table 6. The BO trials against validation loss are plotted in Figure 11. The initial 6 trials were based on Sobol sequencing. The plot shows that convergence occurred within 25 trials. Similar to the MLP, the MDN also seemed to have a relatively flat loss landscape with possible broad minima. This indicates that further hyper-parameter tuning within the current search space will likely not provide much value.

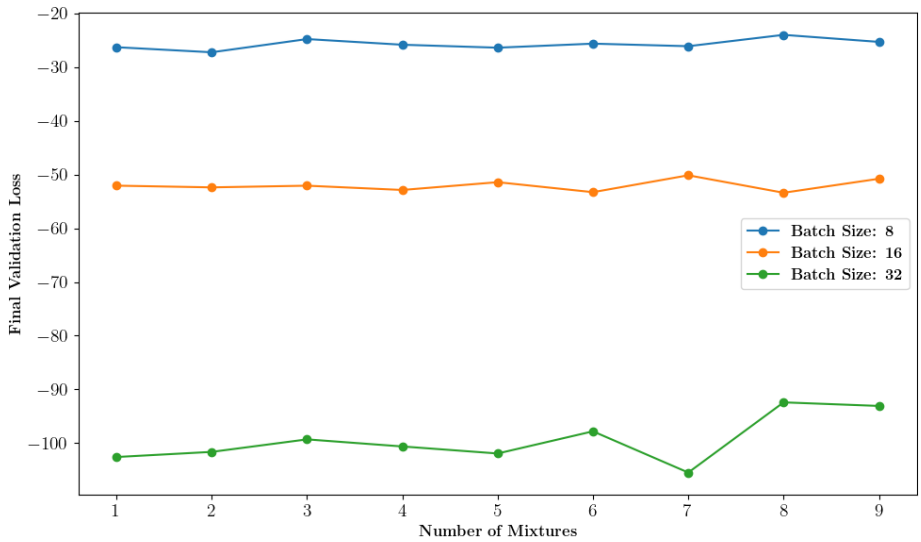


Figure 10. MDN grid search: Number of mixtures for different batch sizes against final validation loss (LR: 1×10^{-5}, Architecture: 32-32-32, Act Function: ELU).

Table 6. Hyper-parameter search space for MDN Bayesian-search.

Hyper-Parameter	**Values**
Neuron count per layer	8, 32, 64
Number of hidden layers	2, 3
Activation function	ELU
Learning rate	Log range from 1×10^{-6} to 1×10^{-3}
Batch Size	32
Number of mixtures	1–7
Determined Hyper-Parameters	**Values**
Neuron count per layer	32
Number of hidden layers	2
Activation function	ELU
Learning rate	3.48×10^{-5}
Batch Size	32
Number of mixtures	4

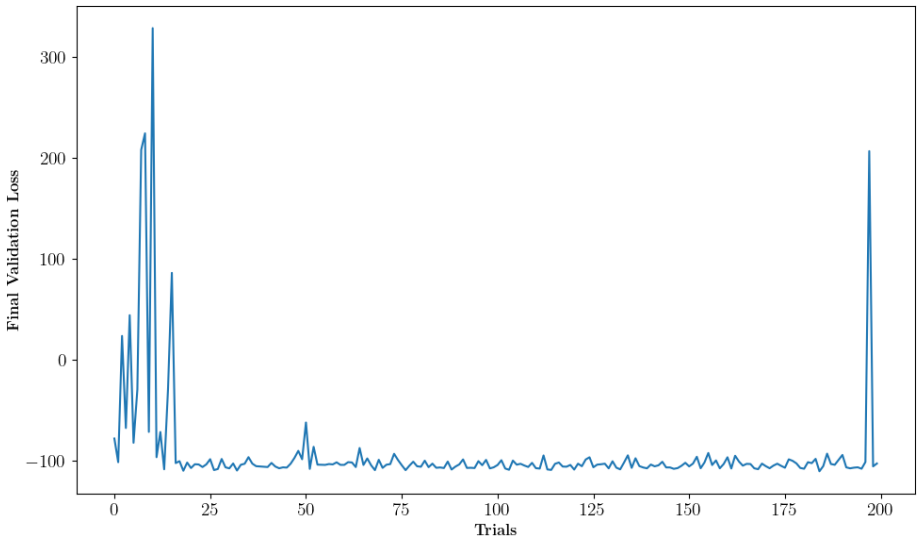


Figure 11. MDN BO: Bayesian optimisation showing trials against final validation loss. Note: two extreme outliers (trial 54 and trial 189) were removed for visualisation purposes.

3.1.4. Random Forest

For the final model, a grid search was not deemed necessary as the random forest model is relatively computationally inexpensive. This allowed a larger, more broad Bayesian search. The 500 BO trials plotted against validation loss are shown in Figure 12. The first 8 trials used Sobol sequencing. Similar to the previous applications of the BO algorithm on the MLP and MDN, the results also suggest that the RF network was mostly insensitive to hyper-parameter changes. The hyper-parameter search space parameters and the hyper-parameter set with the lowest RMSE (Trial 4) are shown in Table 7. As trial 4 had the lowest RMSE value for all tested hyper-parameter sets, this indicates that the final set was found during the initial Sobol sequencing.

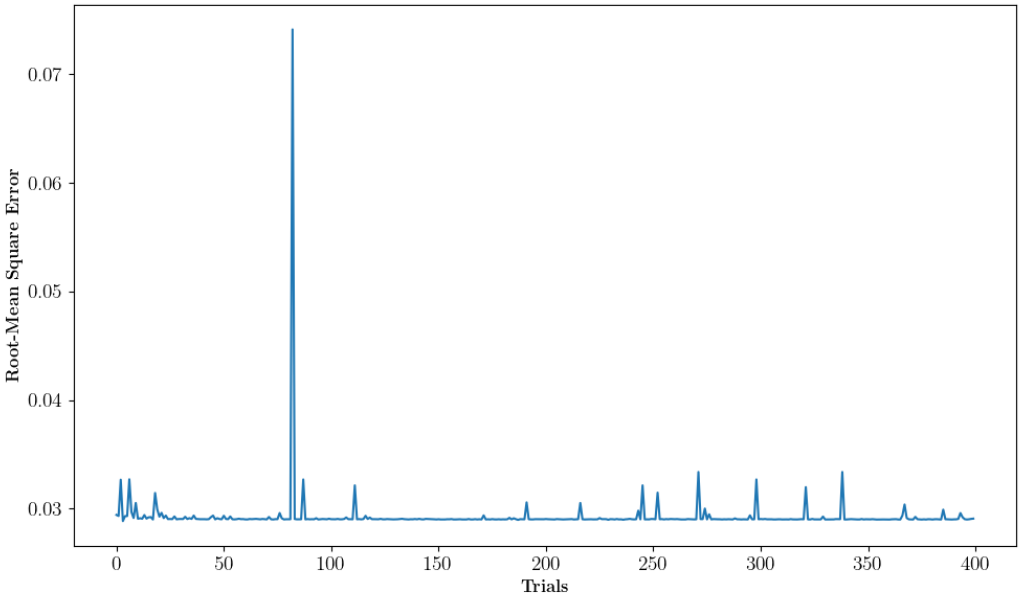


Figure 12. RF BO: Bayesian optimisation showing trials against RMSE.

Table 7. Hyper-parameter search space for RF.

Hyper-Parameter Search Space	**Values**
Number of trees	1:2000
Max Tree Depth	1:2000
Min number of samples per leaf	1:50
Min number of samples per split	2:50
Determined Hyper-Parameters	**Values**
Number of trees	13
Max Tree Depth	680
Min number of samples per leaf	15
Min number of samples per split	23

3.2. Model Comparison

All results in this section were derived from the averaged outputs of 10-fold cross-validation. This method has shown similar or better results compared to leave-one-out cross-validation [36]. Furthermore, all error metrics are defined based on the difference between the estimated lower facial measurements and the true lower facial measurements, as measured in Section 2.1. In comparing the final models' performance on the test data, these errors were also evaluated in relation to the clinical thresholds.

The hyper-parameter optimisation as per Section 3.1.1 was repeated for a modified dataset. The modified dataset, from here on referred to as the symmetry dataset, assumed upper facial symmetry by dropping all upper right-side facial measurements as input

features. Note that with the initial dataset, only lower facial symmetry was assumed. This comparison provides insight into the effect of minor asymmetry and the number of input features on prediction accuracy. The resultant validation mean absolute error (MAE) is shown in Table 8. These results indicate that the inclusion of the right-side input features in the asymmetry dataset provides an overall improvement in the average performance across models. However, there is no noticeable impact on the linear regression model's ability to predict angles. This discrepancy may suggest that angular measurements have a higher complexity than distance measurements. Speculatively, the improvement in the non-linear methods may indicate that, in the presence of more data, these models still have residual learning capacity available to learn meaningful relationships, leading to lower prediction errors. In contrast, it is possible that by adding additional features, the linear regression model did not have the capacity to learn additional complex relationships in the angular target features. Assuming this trend holds in the test data, and given the improved average performance across models in the validation dataset, the analysis of the testing (hold-out) dataset proceeded with the asymmetrical dataset.

Table 8. Change in validation loss with addition of features. "Asymmetric" refers to the dataset with additional features.

	MLP	MDN	RF	LR
	Distances (validation)			
Symmetry: MAE (mm)	2.91	2.9	3.05	2.98
Asymmetry: MAE (mm)	2.7	2.74	2.77	2.78
	Angles (validation)			
Symmetry: MAE (°)	3.21	3.24	3.14	3.01
Asymmetry: MAE (°)	2.93	2.92	2.92	3.01

The mean absolute error (MAE), root-mean-square error (RMSE), and mean average percentage error (MAPE) were calculated for each feature and then averaged across features with similar measurement types, namely distances and angles. The test set MAE values for the distance features with the MLP, MDN, RF, and LR models ranged respectively from 1.85 to 4.00 mm, 1.83 to 4.03 mm, 1.88 to 4.06 mm, and 1.79 to 4.14 mm. Similarly, for angular measurements, the MAEs ranged respectively from 2.27 to 3.75°, 2.25 to 3.75°, 2.08 to 3.78°, and 2.57 to 3.42°. The error metrics from the testing dataset are presented in Table 9. The calculated metrics for the distance measurements indicate comparable error values across models. The MLP and MDN had slightly lower errors. In contrast, for angular measurements, accounting for noise, the MAE values showed almost no differences. Using the more outlier-sensitive RMSE, the results indicate that non-linear methods have barely noticeable lower angle error scores. Additionally, the MAPE values for both measurement types did not show discernible differences between models.

Table 9. Average error metrics for the test (hold-out) dataset.

	MLP	MDN	RF	LR
	Distances (test)			
MAE (mm)	2.77	2.79	2.95	2.91
RMSE (mm)	3.72	3.74	3.85	4.01
MAPE (%)	0.08	0.08	0.08	0.08
	Angles (test)			
MAE (°)	3.09	3.11	3.07	3.12
RMSE (°)	3.73	3.73	3.76	3.84
MAPE (%)	0.04	0.04	0.04	0.04

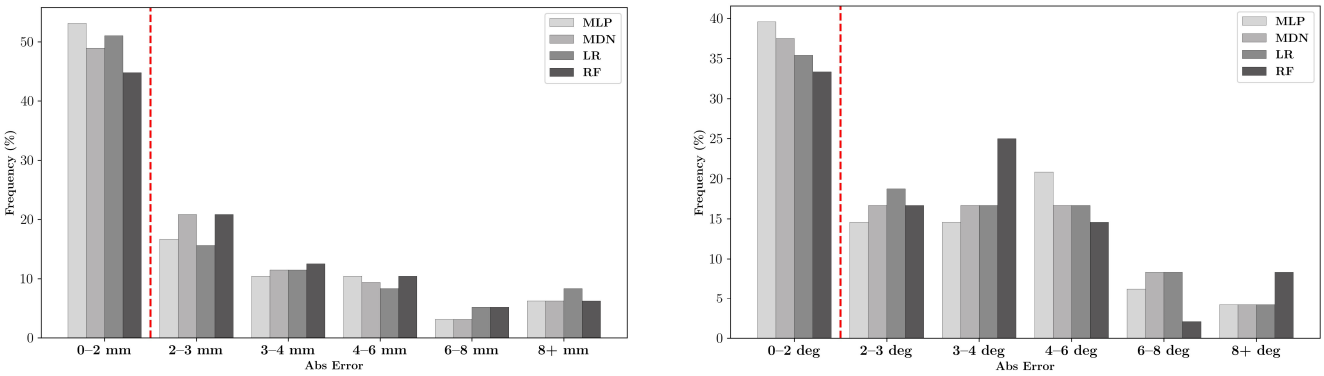
Among the models evaluated, while neural network-based models achieved the lowest scores, none showed noticeably lower error values than any other. A direct comparison with other studies is not feasible due to variations in data dimensions, differences in features (landmarks and different cephalometric measurements), and input feature sets that span both the upper and lower face during prediction. Nevertheless, these studies may provide insight into expected error ranges and possible trends.

The study by Cheng [4], which focused on predicting mandible repositioning vectors, found that their transformer-based model (MAE: 1.34 mm), along with MLP (MAE: 1.36 mm), RF (MAE: 1.41 mm), and ridge regression model (MAE: 1.37 mm), also demonstrated comparable performance. They concluded that their transformer had slightly better generalisation capability on the test data compared to other models. The transformer model's MAE values ranged from 1.08 to 1.80 mm. In Gillingham's research [9], multiple linear regression models were used to predict mandibular cephalometric measurements from other mandibular and upper-facial measurements. Their RMSE values ranged from 1.74° to 5.01° and from 0.64 mm to 2.89 mm. It is important to note that RMSE typically provides estimates that are as conservative or more conservative than those provided by MAE. Finally, Zhu [10] applied a landmark-based approach to predict the mandible from the maxilla and vice versa. In predicting the mandible, they found that their 2-dimensional linear regression performed best, providing average scalar position errors ranging from approximately 2.9 mm to 7.9 mm.

In relation to the current work, the lower errors from Gillingham's model may be attributed to the inclusion of mandible features with better correlation (compared to upper facial measurements) as input features. Similarly, Cheng's model also utilised both upper and lower facial cephalometric measurements as input features. However, as previously mentioned, Cheng predicted vectors containing landmark coordinates and not cephalometric measurements. Another difference between the studies is Cheng's larger sample size (n = 383) as opposed to the current work (n = 155), Gillingham's research (n = 100), and Zhu's sample size (n = 42). In addition, both Gillingham and Cheng applied a more strict exclusion protocol, excluding subjects based on malocclusion, likely resulting in more harmonious relationships between features.

As before, given the clinical requirements (2 mm/°) for virtual surgery planning, it is important to determine the model's performance in relation to this. Therefore, the absolute errors are also represented in frequency diagrams, with distances shown in Figure 13a and angles in Figure 13b. In descending order, for distance errors, the MLP, LR, MDN, and RF models had a relative frequency of errors within the threshold of 53.13%, 51.04%, 48.96%, and 44.79%, respectively. For angular measurements, the MLP, MDN, LR, and RF models had a relative frequency of errors within specifications of 39.58%, 37.5%, 35.42%, and 33.33%, respectively. Notably, if a more lenient cut-off of 3 mm/° is used, then the difference in cumulative frequency of errors between models diminishes. For distances, the MLP, MDN, LR, and RF models had relative frequencies of 69.79%, 69.79%, 66.67%, and 65.62%, respectively. For angles, the new relative error frequencies were 54.17%, 54.17%, 54.17%, and 50.00%, respectively.

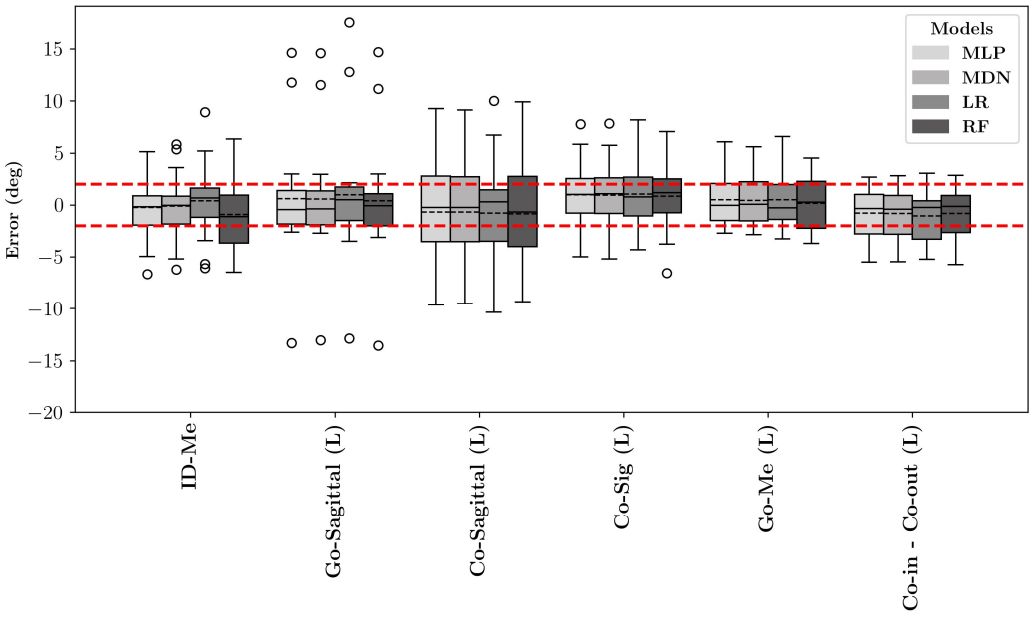
Across models, it is evident that the models have more difficulty learning the relationship of input features to angular target features compared to distance target features. In terms of model-specific performance, in both the lenient and conservative cases (3 and 2 mm/°, respectively), the RF model had the highest frequency of errors outside of specification. Conversely, the MLP had the most errors under the conservative threshold by a slight margin. The MLP and MDN outperform the other models under the lenient criteria. A limitation of the RF model is that it cannot extrapolate outside of its training data [37]. This inability may contribute to the RF model's high error frequency.



(**a**) (**b**)

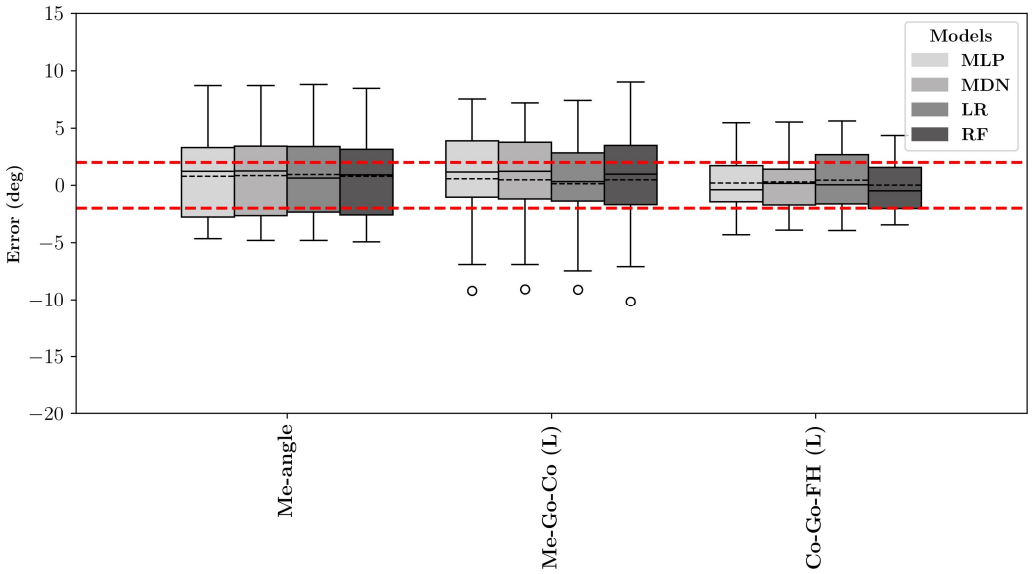
Figure 13. Frequency of errors for predicted testing dataset measurements. The dashed line indicates the clinical specification of 2 mm/°. (**a**) Absolute errors for distances. (**b**) Absolute errors for angles.

The model outputs can be interpreted on a feature basis. Figure 14a,b show the distributions of signed validation errors for each model and feature. This again shows similar error ranges across models, with very few prominent differences. Generally, the signed mean and median values were consistent across both models and features. On a feature basis, the error distributions from the different models tended to have a similar skew. In terms of error outliers, these were predominantly caused by the same individuals and were similar across models.



(**a**)

Figure 14. *Cont.*



(**b**)

Figure 14. Error distribution per feature per model. The dashed red line indicates the clinical specification of 2 mm/°. The solid line in the box represents the median, whereas the dashed line in the box represents the mean. (**a**) Signed error distribution for distances. (**b**) Signed error distribution for angles.

3.3. Implications, Limitations, and Future Directions

In summary, this study provided a unique comparison between linear and non-linear methods for predicting lower facial cephalometric measurements from upper facial cephalometric measurements. Similarly, the combination of methods compared was also unique. An overview comparing key differences between the current work and related research is provided in Table 10. In addition, this study was the first to construct regression models based on cephalometric measurements derived from the publicly available HNSCC dataset. Lastly, to the authors' knowledge, a mixture density network has not previously been applied to cephalometric measurements.

Table 10. An overview of similar research.

Study	General Application	Measurement Based Models	Orthognathic Focus	Coordinate Based Models	Non-Linear Approach	Upper and Lower Facial Separation
Wang et al. [1]	Jaw prediction with upper/lower facial split		x	x		x
Cheng et al. [4]	Surgery plan prediction with dentition, jaw and soft tissue	x	x	x	x	
Segner [7]	Harmony evaluation	x	x			
Bingmer et al. [8]	Harmony evaluation	x	x			
Gillingham et al. [9]	Mandible prediction with partial inputs	x				
Zhu et al. [10]	Mandible/Maxilla prediction with upper/lower facial split			x		x
Current Study	Mandible prediction with upper/lower facial split	x	x		x	x

From the model comparison in Section 3.2, it is apparent that in predicting 1D lower facial measurements from 1D upper facial measurements, non-linear models provide similar or marginally better performance compared to the traditional OLS linear models. Specifically, deep learning-based models showed lower mean absolute errors and better

adherence of errors to clinical specifications than the OLS and RF models. In addition, the analysis showed that all evaluated models frequently produced errors exceeding the clinical threshold of 2 mm/degree. This indicates that refinement is required. Furthermore, the exhaustive hyper-parameter search in Section 3.1 revealed no noticeable difference between 2 and 3 layers for neural network-based models. Although deeper layers are commonly associated with the capability to model more complex relationships [38], this finding suggests that the nonlinear relationship might not be as prominent in the current dataset. This notion is supported by the observation that the OLS model only performed marginally worse compared to the nonlinear models, indicating that the OLS model's inherent inductive bias of assumed linearity may be sufficient for this dataset. As a result, although deep-learning methods slightly outperformed the linear regression model, the added effort involved in their training is difficult to justify based on prediction accuracy alone. However, the information density and noise levels of this study's chosen 1D measurements might be insufficient for the non-linear models examined to effectively learn relationships between the features. Additional data may better describe the relationship between lower and upper facial measurements.

Therefore, one potential limitation of this study is that the current sample size may be too small for non-linear models to effectively learn meaningful non-linear relationships. Furthermore, the current study's exclusion protocol did not exclude individuals with malocclusion. Non-harmonious relationships may have led to poorer model specificity. It is speculated that model performance can be refined by increasing the training sample size and implementing a stricter exclusion protocol. Furthermore, the input and output features were based on manually placed landmarks by a single rater. This approach may introduce bias and noise into the measurements. To mitigate this, it has been proposed that duplicate measurements should be taken and averaged [39] or to use automatic landmarking algorithms [40]. In this project, to assess rater noise, ICC tests were used and found agreement scores ranging between 0.7 and 1.0, with an average of 0.95, indicating that landmark-placement repeatability ranged from acceptable (>0.7) to excellent (>0.9) [41]. In addition, to reduce variability in predictions, 10-fold cross-validation was employed on the hold-out set. Finally, the resulting models are only applicable to the demographics of the training set. Research has shown statistically significant differences in cephalometric measurements for different population groups [42].

In terms of clinical applications, in predicting lower facial measurements from upper facial measurements, the MLP was generally the best for both conservative and lenient tolerances (respectively, 2 mm/° and 3 mm/°). However, for lenient tolerances, the differences between the MLP and MDN were marginal, and the MDN has the added advantage of providing an estimate of the prediction's uncertainty. In addition, Zhu et al. [10] recommended that models should ideally provide the clinician with a selection of estimated values. An MDN can fulfil this criterion by enabling the clinician to sample estimated measurements from the predicted distribution. The clinician can then select their preferred value based on their experience and the patient's requirements.

For future work, these results indicate that non-linear relationships between upper and lower facial measurements may be present but are not significant or well-captured by the current selection of models. It is also possible that the non-linearity in the current dataset is dominated by noise. However, the non-linear relationship could become more apparent at higher dimensions. Shape models such as SSMs, which are essentially linear regression models that employ Principal Component Analyses to reduce extremely dense features to a limited parameter set, have shown improved performance compared to 1D linear regression models in predicting derived measurements [9]. Therefore, it is proposed that further investigation is required in utilising linear shape models and comparing them to non-linear shape models for the prediction of lower facial shape based on upper facial shape. Shape models can also be used to generate artificial data to provide 1D measurement-based models with additional data. For future work regarding 1D measurement models, further comparison should consider the use of regularisation methods for neural network-based

models. In a general application of neural networks, evidence indicates that the application of regularisation "cocktails" may allow neural-network-based methods to outperform state-of-the-art tabular data models [43]. It is possible that this applies to cephalometric measurement datasets. It has also been suggested that neural networks may have improved performance if uninformative features are removed [44]. Therefore, due to the suspected high presence of multi-collinearity between input measurements in this current dataset, along with the ICC scores indicating that certain feature definitions may be more susceptible to measurement error or noise than others, it is recommended that feature selection should be considered.

4. Conclusions

This study aimed to determine whether non-linear models outperform linear models in predicting lower facial measurements from upper facial measurements. Accordingly, linear regression, random forest, multilayer perceptron, and mixture density network models were implemented. The neural networks and the random forest underwent an exhaustive grid and Bayesian search for hyper-parameter optimisation, whereas feature-space transforms were applied to the linear regression. The validation and testing results indicated that all models performed at comparable levels; however, the neural network-based models had marginally more errors below clinically relevant thresholds (2 mm/°) and lower MAE values than the linear regression and random forest models. Therefore, based on the results of this study, an MDN is suggested due to its ability to produce estimates of uncertainty along with predicted values. Nevertheless, future work is required to further improve prediction accuracy under conservative thresholds.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/mca29040061/s1, File S1: HNSSC_subjects.

Author Contributions: Conceptualization, J.T., J.v.d.M. and R.L.; methodology, J.T., J.v.d.M. and R.L.; software, J.T. and J.v.d.M.; validation, J.T.; formal analysis, J.T.; investigation, J.T.; resources, J.T., J.v.d.M. and R.L.; data curation, J.T. and J.v.d.M.; writing—original draft preparation, J.T.; writing—review and editing, J.T., J.v.d.M. and R.L.; visualization, J.T.; supervision, J.v.d.M. and R.L.; project administration, J.T. and J.v.d.M.; funding acquisition, J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data. Raw data were obtained from the HNSCC repository (https://doi.org/10.7937/k9/tcia.2020.a8sh-7363, accessed on 20 July 2023) managed by The Cancer Imaging Archive (TCIA). Processed data are available from the authors with permission obtained from the TCIA.

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

Appendix A

Table A1. Definitions of mandible landmarks.

Landmark	Definition
Menton	Most inferior point on symphysis menti located on sagittal plane
Condylion	Most superior point on condylar process
Sigmoid notch	Deepest point between the condylar and coronoid process
Coronoid process apex	Most superior point on coronoid process
Gonion	Most inferior, lateral and posterior point on the external angle
Infradentale	Most anterior point on mandible alveolar process located on sagittal plane
Pogonion	Most anterior point on symphysis menti located on sagittal plane
B point	Deepest point between Id and Pg located on sagittal plane
Gnathion	Median point between Pg and Me located on sagittal plane
Lateral condyle	Most lateral point on the condylar process
Medial condyle	Most medial point on the condylar process

Table A2. Definitions of cranium and maxilla landmarks.

Landmark	Definition
Frontomalare-orbitale	Lateral point on orbit rim located at distinct change of curvature where zyogomatic bone meets with the frontal bone
Orbitale	Most inferior point on lower orbit rim
Maxillary point	Most concave point on zygomatic process of the maxilla
Porion	Most superior point on the upper margin of the external auditory meatus
Nasion	Deepest point on concavity formed by nasal bone and frontal bone
Anterior nasal spine	Most anterior point of nasal floor projecting from maxilla
Prosthion	Most anterior point on lower maxilla located on sagittal plane
A point	Deepest point between ANS and Prosthion located on midsagital plane
Sella	Most inferior point on sella turcica located on midsagittal plane
Posterior nasal spine	Most posterior point of nasal floor projecting from maxilla

References

1. Wang, L.; Ren, Y.; Gao, Y.; Tang, Z.; Chen, K.-C.; Li, J.; Shen, S.G.F.; Yan, J.; Lee, P.K.M.; Chow, B.; et al. Estimating patient-specific and anatomically correct reference model for craniomaxillofacial deformity via sparse representation. *Med. Phys.* **2015**, *42*, 5809–5816. [CrossRef] [PubMed]
2. Dall'Magro, A.K.; Dogenski, L.C.; Dall'Magro, E.; Figur, N.S.; Trentin, M.S.; De Carli, J.P. Orthognathic surgery and orthodontics associated with orofacial harmonization: Case report. *Int. J. Surg. Case Rep.* **2021**, *83*, 106013. [CrossRef] [PubMed]
3. Murphy, C.; Kearns, G.; Sleeman, D.; Cronin, M.; Allen, P.F. The clinical relevance of orthognathic surgery on quality of life. *Int. J. Oral Maxillofac. Surg.* **2011**, *40*, 926–930. [CrossRef] [PubMed]
4. Cheng, M.; Zhang, X.; Wang, J.; Yang, Y.; Li, M.; Zhao, H.; Huang, J.; Zhang, C.; Qian, D.; Yu, H. Prediction of orthognathic surgery plan from 3D cephalometric analysis via deep learning. *BMC Oral Health* **2023**, *23*, 161. [CrossRef] [PubMed]
5. Hammoudeh, J.A.; Howell, L.K.; Boutros, S.; Scott, M.A.; Urata, M.M. Current Status of Surgical Planning for Orthognathic Surgery: Traditional Methods versus 3D Surgical Planning. *Plast. Reconstr. Surg. Glob. Open* **2015**, *3*, e307. [CrossRef] [PubMed]
6. Reyneke, J.P.; Ferretti, C. Diagnosis and Planning in Orthognathic Surgery. In *Oral and Maxillofacial Surgery for the Clinician*; Bonanthaya, K., Panneerselvam, E., Manuel, S., Kumar, V.V., Rai, A., Eds.; Springer Nature: Singapore, 2021; pp. 1437–1462. ISBN 9789811513466.
7. Segner, D. Floating norms as a means to describe individual skeletal patterns. *Eur. J. Orthod.* **1989**, *11*, 214–220. [CrossRef] [PubMed]
8. Bingmer, M.; Özkan, V.; Jo, J.; Lee, K.-J.; Baik, H.-S.; Schneider, G. A new concept for the cephalometric evaluation of craniofacial patterns (multiharmony). *Eur. J. Orthod.* **2010**, *32*, 645–654. [CrossRef] [PubMed]
9. Gillingham, R.L.; Mutsvangwa, T.E.M.; van der Merwe, J. Reconstruction of the mandible from partial inputs for virtual surgery planning. *Med. Eng. Phys.* **2023**, *111*, 103934. [CrossRef] [PubMed]
10. Zhu, X.; Han, J.; Zhang, S.; Min, X.; Liu, J.; Zhai, G. Restoring Skeletal Marker Points for Severe Maxillary and Mandibular Jaw Defects Using a Linear Regression Approach. *J. Oral Maxillofac. Surg.* **2019**, *77*, 664.e1–664.e16. [CrossRef]
11. Gillingham, R.L.; Delport, S.; Van Der Merwe, J.; Mutsvangwa, T. Retrospective study on mandibular morphology towards aiding implant design. In Proceedings of the 2018 3rd Biennial South African Biomedical Engineering Conference (SAIBMEC), Stellenbosch, South Africa, 4–6 April 2018; pp. 1–4.
12. Alkhayer, A.; Piffkó, J.; Lippold, C.; Segatto, E. Accuracy of virtual planning in orthognathic surgery: A systematic review. *Head Face Med.* **2020**, *16*, 34. [CrossRef] [PubMed]
13. M.D. Anderson Cancer Center Head and Neck Quantitative Imaging Working Group; Grossberg, A.; Elhalawani, H.; Mohamed, A.; Mulder, S.; Williams, B.; White, A.L.; Zafereo, J.; Wong, A.J.; Berends, J.E.; et al. *HNSCC Version 4*. [Dataset]. The Cancer Imaging Archive. Available online: https://doi.org/10.7937/k9/tcia.2020.a8sh-7363 (accessed on 20 July 2023).
14. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.-C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **2012**, *30*, 1323–1341. [CrossRef] [PubMed]
15. Bayome, M.; Park, J.H.; Kook, Y.-A. New three-dimensional cephalometric analyses among adults with a skeletal Class I pattern and normal occlusion. *Korean J. Orthod.* **2013**, *43*, 62–73. [CrossRef] [PubMed]
16. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]
17. Liljequist, D.; Elfving, B.; Skavberg Roaldsen, K. Intraclass correlation—A discussion and demonstration of basic features. *PLoS ONE* **2019**, *14*, e0219854, S1 Appendix: ANOVA. Sum of squares and mean squares–definitions and relations. [CrossRef] [PubMed]
18. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
19. 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. *arXiv* **2019**. [CrossRef]

20. Balandat, M.; Karrer, B.; Jiang, D.R.; Daulton, S.; Letham, B.; Wilson, A.G.; Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. *arXiv* **2020**. [CrossRef]
21. Bakshy, E.; Dworkin, L.; Karrer, B.; Kashin, K.; Letham, B.; Murthy, A.; Singh, S. AE: A domain-agnostic platform for adaptive experimentation. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, QC, Canada, 3–8 December 2018.
22. Mehta, P.; Bukov, M.; Wang, C.-H.; Day, A.G.R.; Richardson, C.; Fisher, C.K.; Schwab, D.J. A high-bias, low-variance introduction to Machine Learning for physicists. *Phys. Rep.* **2019**, *810*, 1–124. [CrossRef] [PubMed]
23. Bishop, C.M. *Pattern Recognition and Machine Learning*; Information science and statistics; Springer: New York, NY, USA, 2006; ISBN 978-0-387-31073-2.
24. Abu-Mostafa, Y.S.; Magdon-Ismail, M.; Lin, H.-T. Learning from Data. Available online: https://amlbook.com/index.html (accessed on 14 June 2024).
25. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **1989**, *2*, 303–314. [CrossRef]
26. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
27. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**. [CrossRef]
28. Bishop, C.M. *Mixture Density Networks*; Aston University: Birmingham, UK, 1994; [Technical Report].
29. He, L.; Levine, R.A.; Fan, J.; Beemer, J.; Stronach, J. Random Forest as a Predictive Analytics Alternative to Regression in Institutional Research. *Pract. Assess. Res. Eval.* **2019**, *23*, 1.
30. Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Chapman and Hall/CRC: New York, NY, USA, 2017; ISBN 978-1-315-13947-0.
31. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
32. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; Freitas, N.D. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **2016**, *104*, 148–175. [CrossRef]
33. Sobol', I.M. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 86–112. [CrossRef]
34. Liu, R.; Gillies, D.F. Overfitting in linear feature extraction for classification of high-dimensional image data. *Pattern Recognit.* **2016**, *53*, 73–86. [CrossRef]
35. Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv* **2016**. [CrossRef]
36. Raghu, M.; Poole, B.; Kleinberg, J.; Ganguli, S.; Sohl-Dickstein, J. On the Expressive Power of Deep Neural Networks. *arXiv* **2017**. [CrossRef]
37. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2, Montreal, QC, Canada, 20–25 August 1995; pp. 1137–1143.
38. Hengl, T.; Nussbaum, M.; Wright, M.; Heuvelink, G.; Graeler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* **2018**, *6*, e5518. [CrossRef]
39. Houston, W.J.B. The analysis of errors in orthodontic measurements. *Am. J. Orthod.* **1983**, *83*, 382–390. [CrossRef] [PubMed]
40. Junaid, N.; Khan, N.; Ahmed, N.; Abbasi, M.S.; Das, G.; Maqsood, A.; Ahmed, A.R.; Marya, A.; Alam, M.K.; Heboyan, A. Development, Application, and Performance of Artificial Intelligence in Cephalometric Landmark Identification and Diagnosis: A Systematic Review. *Healthcare* **2022**, *10*, 2454. [CrossRef] [PubMed]
41. Goulart, M.E.P.; Biegelmeyer, T.C.; Moreira-Souza, L.; Adami, C.R.; Deon, F.; Flores, I.L.; Gamba, T.O. What is the accuracy of the surgical guide in the planning of orthognathic surgeries? A systematic review. *Med. Oral Patol. Oral Cir. Bucal* **2022**, *27*, e125–e134. [CrossRef] [PubMed]
42. Baruah, N.; Bora, M. Cephalometric Evaluation Based on Steiner's Analysis on Young Adults of Assam. *J. Indian Orthod. Soc.* **2009**, *43*, 17–22. [CrossRef]
43. Kadra, A.; Lindauer, M.; Hutter, F.; Grabocka, J. Well-tuned Simple Nets Excel on Tabular Datasets. *arXiv* **2021**. [CrossRef]
44. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? *arXiv* **2022**. [CrossRef]