Pattern Recognition Letters 195 (2025) 16–22

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Graph Neural Networks for 3D facial morphology: Assessing the effectiveness of anthropometric and automated landmark detection

Giuseppe Maurizio Facchi [a],*, Giuliano Grossi [a], Alessandro D’Amelio [a], Francesco Agnelli [a], Chiarella Sforza [b], Gianluca Martino Tartaglia [c,d], Raffaella Lanzarotti [a]

[a] PHuSe Lab, Department of Computer Science, University of Milan, Milan, Italy
[b] LAFAS - Department of Biomedical Sciences for Health, University of Milan, Milan, Italy
[c] Department of Biomedical, Surgical and Dental Sciences University of Milan, Milan, Italy
[d] Ospedale Maggiore Policlinico, UOC Maxillo-Facial Surgery and Dentistry Fondazione IRCCS Cà Granda, Milan, Italy

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Graph Neural Networks
3D
Morphological features
Face
Facial dimorphisms
Facescape
Dad3dheads

ABSTRACT

This study investigates the potential of Graph Neural Networks (GNNs) for analyzing 3D facial morphology, leveraging facial landmarks as graph nodes to capture the intrinsic structure of 3D face scans. This research evaluates the effectiveness of three distinct approaches for defining graph vertices by associating them with: (1) a well-established set of anthropometric landmarks identified through tactile assessment, widely considered the gold standard in facial anthropometry; (2) automatically detected 3D facial keypoints estimated using advanced algorithms; and (3) geometry-based random point cloud sub-sampling via farthest point sampling (FPS). To evaluate the effectiveness of GNNs and facial landmarks in capturing and representing meaningful morphological patterns, the study employs two benchmark tasks: gender classification and age regression. Extensive experiments across various GNN architectures and three datasets — each presenting diverse and challenging conditions — demonstrate that semantically meaningful landmarks, whether anthropometric or automatically detected, consistently outperform non-semantic random samples in both tasks and across all datasets. These results highlight the crucial role of semantic contextualization in graph-based facial analysis. Notably, models utilizing automatically detected facial keypoints achieved performance comparable to those based on manually annotated anthropometric landmarks, offering a scalable and cost-effective alternative without compromising accuracy. These findings support the integration of automated GNN-based methodologies into a wide range of applications, including clinical diagnosis, forensic analysis, and biometric recognition.

1. Introduction

Beyond its fundamental role in perception, the human face acts as a rich medium for conveying diverse information [1], including biometric traits such as identity, gender, age, and ethnicity [2], as well as physiological [3] and behavioral responses [4]. Additionally, the face offers valuable cues about clinical status [5]. Notably, many genetic syndromes can lead to craniofacial abnormalities [6,7]. Consequently, analyzing facial morphology is particularly relevant in various fields, including the identification of craniofacial pathologies, cosmetic surgery, protective gear design, and surgical reconstruction, among others [5].

In medical literature, the use of facial anthropometry based on manually detected landmarks is well-documented [8]. Once identified, these landmarks are used to characterize the facial morphological features associated with the phenotype under investigation. As shown in [9], even basic facial measurements, such as linear distances and angles between landmarks, provide an effective approach for facial analysis. However, manual landmark detection is highly demanding, requiring precise execution by trained experts. The time required for identifying reference points, performing photometric stereo analysis, and completing patient discharge makes this technique economically impractical for confirming a physician’s diagnostic suspicion.

---

* Corresponding author.

E-mail addresses: giuseppe.facchi@unimi.it (G.M. Facchi), giuliano.grossi@unimi.it (G. Grossi), alessandro.damelio@unimi.it (A. D’Amelio), francesco.agnelli@unimi.it (F. Agnelli), chiarella.sforza@unimi.it (C. Sforza), gianluca.tartaglia@unimi.it (G.M. Tartaglia), raffaella.lanzarotti@unimi.it (R. Lanzarotti).

https://doi.org/10.1016/j.patrec.2025.04.028
Received 8 January 2024; Received in revised form 10 March 2025; Accepted 24 April 2025
Available online 9 May 2025

This study explores the automatic identification of landmarks in 3D face scans. Validating this approach could offer a scalable and broadly applicable solution for supporting medical diagnosis, enabling remote supervision of the diagnostic process. This, in turn, could enhance the efficiency of large-scale population screening and improve access to diagnostic services.

The extraction of meaningful morphological patterns is entrusted to Graph Neural Network (GNN) architectures, which are powerful tools for representation learning in graphs [10]. GNNs have been successfully applied across various domains, including computer vision, natural language processing, chemistry, physics, biology, traffic networks, and recommendation systems [11]. Despite the rapid proliferation of GNN models [10,12–14], incorporating diverse propagation, sampling, and pooling mechanisms [11], relatively little attention has been given to defining the graph structure for optimal predictive performance. This oversight is largely due to the early adoption of GNNs in domains with inherently defined graph structures, such as molecular modeling or physical systems, where structural relationships are explicitly provided, simplifying their application. In contrast, in domains lacking predefined graph structures — where the graph must be constructed or learned from data — additional effort is required to establish an appropriate representation [15–17]. In this context, the study of 3D faces presents a hybrid case [18]. Unlike molecules, 3D faces lack a straightforward structure. However, based on their configuration, they can be robustly described by a set of facial anthropometric landmarks, which serve as nodes in the graph [10].

This paper investigates the impact of automated vertex detection on the ability of GNN models to effectively capture and represent facial morphological features. Specifically, we focus on facial point clouds and compare the performance of three distinct vertex characterization approaches. The first approach relies on anthropometric landmarks manually identified by experts in facial morphology. The second utilizes facial landmarks automatically generated by the MVLM model proposed by Paulsen et al. [19]. Finally, we examine a more flexible vertex definition using the farthest point sampling (FPS) technique, which selects points without any specific semantic association.

To ensure robust conclusions independent of specific methodologies and configurations, we evaluate multiple GNN models across different connectivity levels. Furthermore, a consistent and comprehensive characterization of nodes and edges is applied across all models. This study examines craniofacial patterns that differentiate gender (sexual dimorphism) and age among subjects [20,21]. However, these tasks serve as a proof-of-concept, highlighting the potential for broader applications, including syndrome diagnosis and identity recognition. Briefly, our objective is to address the following inquiries across multiple models:

RQ1. Does the reliance on semantic nodes or FPS make a difference?
RQ2. Can automatic landmarks serve as a viable alternative to anthropometric landmarks within the context of semantic landmarking?

The remainder of the paper is structured as follows. Section 2 outlines the policy of vertices and edges selection, describes the node and edge attributes, and provides an overview of the tested GNN models. Section 3 presents and discusses the obtained results. Finally, Section 4 summarizes the conclusions.

2. Graph structure and GNN models

In this study, we represent a graph as a pair $G = (V, E)$, where V is the set of nodes and $E \subset V \times V$ is the set of edges. Each node $v_i \in V$ and each edge $e_{ij} = (v_i, v_j) \in E$ connect two nodes v_i and v_j. The neighborhood of a node v_i is defined as $\mathcal{N}(v_i) = \{v_j \in V \mid (v_i, v_j) \in E\}$. Edges are constructed using the k-nearest neighbors (kNN) algorithm, connecting each node to its k closest neighbors. Nodes correspond to facial landmarks, identified through three different methods (see Section 2.1). Each node v is associated with a feature vector $h(v) \in \mathbb{R}^d$, while each edge e_{ij} has an attribute vector $g(e_{ij}) \in \mathbb{R}^c$ (see Section 2.2). This study is conducted in the context of biometric applications, leveraging state-of-the-art graph neural network (GNN) models from the literature (see Section 2.3).

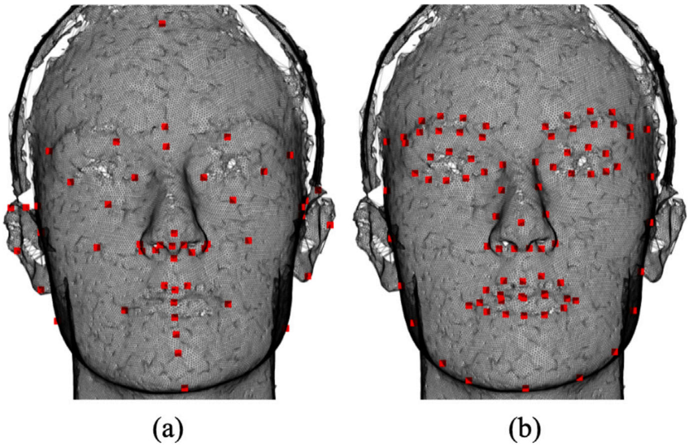

Fig. 1. (a) The set of 50 anthropometric facial landmarks selected by expert anthropometrists via tactile assessment. (b) The set of 84 automatically detected facial landmarks from the MVLM model.

2.1. Landmark detection

The *manual landmark detection* is obtained by an expert who identifies a set of 50 soft-tissue facial anthropometric landmarks through visual inspection or palpation (Fig. 1.a). The identification is based on international criteria and follows an experimental protocol developed and utilized to investigate facial morphometric features, including those associated with genetic syndromes. Landmarks were marked on the cutaneous surface using a common black liquid eyeliner, and then digitized off-line on the 3D facial reconstruction, allowing the x, y, z coordinates of the previously marked anthropometric landmarks to be extracted and the data subsequently elaborated. This approach has the advantage to identify with high precision the facial anthropometric landmarks, but it is invasive, requiring to mark the cutaneous surface, and it is time consuming.

Alternatively, *automatic facial landmark detection* could be employed. There exists a significant literature about 3D Facial landmark detection [22], with methods falling into two main categories: those based on 3D geometric information [23], those relying on statistically learned models [24,25] and deep-learning based methods [26,27]. Here, we adopt a model from the latter category; more precisely the MVLM approach proposed by Paulsen et al. [19]. Briefly, MVLM exploits the 3D face mesh to render several views from different view points and uses a CNN based model to estimate the 2D location of each landmark from each view point. These estimates are then combined using a LSQ and RANSAC approach to have a robust and reliable estimate of the 3D location of 84 landmarks. As far as we know, the MVLM model is the only 3D-face landmarking method deployed in several pretrained versions with a combination of different datasets and rendering methods, which include RGB renderings as well as depth and geometry ones. In this work we refer to the version trained using geometry+depth image channels and with the BU-3DFE dataset, which contains 100 subjects (56 female, 44 male). The method gained an average error of 2.42 mm on localizing 84 3D landmarks (Fig. 1.b).

As an alternative approach, we investigate the effect of using a *uniform sub-sampling* to decrease the cardinality of the initial point cloud to a subset of points retaining no explicit semantic meaning. The FPS technique achieves this by iteratively selecting points from the initial set in a way that each new point chosen is as far away as possible from the ones already selected. This process continues until the desired number of points is reached (for this paper, we set the cardinality to 128). FPS is largely employed in applications concerning point cloud processing [28], 3D object reconstruction [29], and mesh generation [30]. It represents a general purpose technique that guarantees a uniform distribution and effectively captures the fundamental characteristics of the data, while no semantic is attributed to the selected points, and no exact correspondence among different point clouds can be guaranteed.

2.2. Feature extraction and space embedding

Each node is characterized by features that capture different aspects of its properties. First, we consider the *3D coordinates* of each landmark $v_i \in V$, denoted as $\mathrm{pos}(v_i) = (x_i, y_i, z_i)$, which define the spatial position of the node.

Building on this, we derive *geodesic distances*, computed using a heat equation-based method [31]. For each node v_i, we define $\mathrm{geo}(v_i)$ as the collection of geodesic distances between v_i and all its neighbors:

$$\mathrm{geo}(v_i) = \{\mathrm{geo}(v_i, v_j) \mid v_j \in \mathcal{N}(v_i)\}.$$

To further distinguish nodes, we introduce the one-hot encoding of landmark IDs, denoted as $\mathrm{ohe}(v_i)$ [32]. The rationale for incorporating $\mathrm{ohe}(v_i)$ is to enable the model to differentiate landmarks that would otherwise share similar features, such as symmetric or closely positioned landmarks. Since each node in the graph carries distinct semantic information, assigning a unique identifier through one-hot encoding ensures that structurally similar yet semantically different landmarks remain distinguishable. This enhances the model's ability to learn meaningful facial representations.

Finally, to describe the local surface geometry, we compute *Fast Point Feature Histograms* (FPFH) [33]. FPFH captures geometric relationships in a 3D point cloud beyond surface normals and curvature estimations, encoding the mean curvature around each point. This feature representation is invariant to pose (rotations and translations) and robust to variations in sampling density and sensor noise. We denote the FPFH descriptor of each landmark v_i as $\mathrm{FPFH}(v_i)$.

The collected features are arranged in two different configuration embeddings. In the first, all features are attributed to nodes, and no feature is assigned to the edges. Specifically, node embedding $h_1(v)$ involves the horizontal concatenation of four vectors, i.e., point coordinates $\mathrm{pos}(v_i)$, geodesic distances $\mathrm{geo}(v_i)$, one-hot-encoding $\mathrm{ohe}(v_i)$ and histograms $\mathrm{FPFH}(v_i)$, yielding the vector:

$$h_1(v_i) = [\mathrm{pos}(v_i), \mathrm{FPFH}(v_i), \mathrm{geo}(v_i), \mathrm{ohe}(v_i)]. \tag{1}$$

In the second configuration node features are obtained using node-related descriptors, while $\mathrm{geo}(v_i, v_j)$ is used for edge characterization, resulting in the vector:

$$h_2(v_i) = [\mathrm{pos}(v_i), \mathrm{FPFH}(v_i), \mathrm{ohe}(v_i)] \tag{2}$$

for all nodes v_i, and the vector $g(e_{ij})$ for all edges e_{ij}, i.e.,

$$g(e_{ij}) = [\mathrm{geo}(v_i, v_j)]. \tag{3}$$

2.3. GNN models for biometric tasks

We conducted our analysis on two biometric tasks: binary gender classification and age regression. To compare performance, we evaluated five GNN architectures, briefly outlined here.[1]

Graph transformer [34]. For the Graph Transformer model, a two-layer architecture was employed, consisting of 64 and 32 channels in the first and second layers, respectively. Each layer incorporated a dropout rate of 0.2 and was followed by a normalization layer. The output of these layers was processed through a fully connected layer, which took as input the concatenation of three pooling operations — max, mean, and sum — applied to the input.

Graph attention network (GAT) [35]. Similarly, a two-layer architecture with the same number of channels per layer was utilized for the GAT model. The output was passed through a max-pooling layer before being processed by a fully connected layer to produce the final result. Additionally, we evaluated two state-of-the-art convolutional models specifically designed for point-cloud-based data:

DynamicEdgeConv [36]. The architecture for DynamicEdgeConv includes four convolutional layers with dimensions of 64, 64, 128, and 256, respectively. Each layer is followed by batch normalization and a LeakyReLU activation function. After the convolutional layers, a three-layer MLP with 512 and 256 hidden channels was used to perform the final classification. This MLP operates on the concatenation of the global add-pooling and mean-pooling results computed from the input.

PointTransformerConv [37]. The architecture for PointTransformerConv consists of three convolutional layers with dimensions of 16, 32, and 64, respectively. Each layer is preceded by a multi-layer perceptron (MLP) to adjust the input dimensions. A global add-pooling operation is then applied, followed by a two-layer MLP with 64 hidden channels to produce the final logits.

Geometric coordinates neural network (GCNN). Is an architecture whose convolutional layers draw inspiration from a PointNet layer [38], placing particular emphasis on the positional features (see the supplementary material for more information). It is composed by four layers, the first being followed by a Batch Normalization layer. Dropout with a dropout rate of 0.4 is introduced after the ReLU activation function. The final prediction is performed by a fully connected layer after performing max-pooling.

3. Experiments and results

3.1. Datasets

In the experimental analysis, we utilized three distinct datasets that differ in acquisition processes, 3D mesh representations, and facial characteristics. These datasets vary in terms of ethnicity, age, and mesh resolution.

Facial dismorphism dataset (FDD). The FDD dataset was acquired in our laboratory using the Vectra M3 3D imaging system. It includes 318 subjects, comprising 187 men and 131 women, many of whom exhibit some degree of facial dysmorphism. Each subject is represented by a textured 3D mesh. Ages range from 4 to 71 years, with an average age of 20 years. Facial expressions are predominantly neutral. The FDD dataset provides annotations for 50 manually selected anthropometric landmarks [9]. Each landmark was placed on the subject's face by an expert anthropometrist after tactile assessment.

Facescape (FS) dataset [39]. This dataset consists of textured 3D facial models from 847 asian subjects. For each sample, information on gender and 20 distinct facial conditions is provided. Specifically, the dataset includes four primary facial expressions (neutral, sadness, smile, and anger) as well as additional data representing basic facial movements (e.g., lip roll, brow raiser, mouth stretch). Each 3D model contains approximately 25,000 vertices and 50,000 triangles.

DAD-3Dheads (DAD-3DH) dataset [40]. This dataset consists of 44,898 2D images captured in uncontrolled conditions. 42,152 of these images are linked to a FLAME [41] 3D mesh containing 5023 vertices. Each sample is provided with the corresponding gender label.

---

[1] For a more detailed description, refer to the supplementary material.

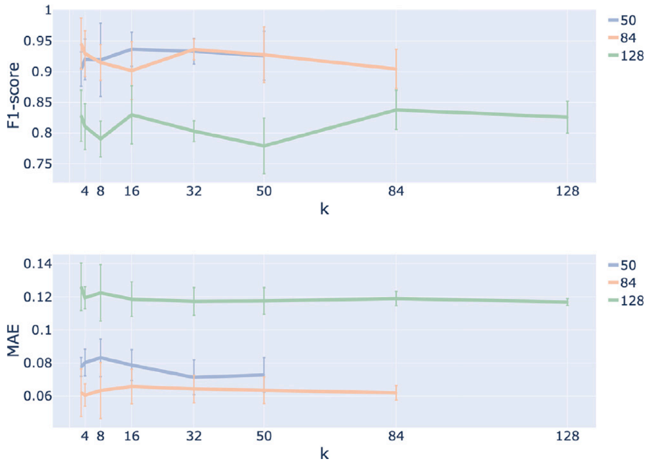

Fig. 2. F1 scores for gender classification (top) and Mean Absolute Errors (MAE) for age prediction (bottom) using the PointTransformerConv network, evaluated across different graph connectivity levels and landmark sets.

3.2. Implementation details

The graph structures vary depending on the node definition and cardinality $(50, 84, 128)$, and the connectivity parameter k in kNN, where k is chosen from $\{3, 4, 8, 16, 32, 50, 84, 128\}$. Since k cannot exceed the total number of vertices, values from 84 onward apply only in specific graph configurations.

During training, we utilized Cross Entropy Loss for the gender classification task and Mean Squared Error Loss for the age prediction task. The models were trained for 100 epochs with a batch size of 32, using Adam optimizer with a learning rate of 10^{-3}. To address data imbalance, the F1 score was reported for gender classification, while the Mean Absolute Error (MAE) metric was used for age prediction. We employed 5-fold cross-validation to assess model performance across all GNN configurations. For convolutional models that support the use of edge attributes, the features described in Eqs. (2) and (3) were utilized (specifically for the GAT and Transformer models). For other models, the feature configuration described in Eq. (1) was applied.

3.3. Results

An initial set of extended experiments was conducted on the FDD dataset, varying connectivity levels and node definitions. The results for each model and task are presented in Tables 1 and 2. As it can be observed, for manually selected landmarks, the PointTransformerConv model consistently outperforms other models across all connectivity configurations. In general, higher connectivity values lead to better performances across experiments. Overall, DynamicEdgeConv and PointTransformerConv demonstrate the best average performance, surpassing other models across all connectivity settings. Fig. 2 illustrates the F1 scores for the gender classification task and the mean absolute error (MAE) for the age estimation task, both obtained using the PointTransformerConv model.

Furthermore, to assess the robustness and generalizability of the proposed approach, we extended our analysis to the Facescape and DAD-3DH datasets, which encompass variations in race, age, and facial expressions. Since these datasets do not provide the set of manually selected anthropometric landmarks, we employed a semi-automatic procedure to select the same set of facial keypoints on these 3D facial scans. The same experts who annotated the FDD dataset manually annotated one face from each dataset. In this case, tactile assessment was infeasible; thus, the anthropometrists relied solely on 3D appearance data. Subsequently, the datasets' 3D morphable face models (which ensure consistent indexing of selected keypoints across all samples) were used to apply the same landmarks to every face. Conversely, the 84 and 128 landmark sets were extracted using the same approach applied to the FDD dataset.

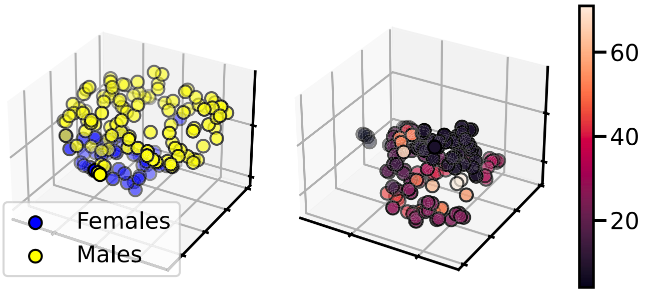

Fig. 3. UMAP 3D projection of the representation learnt by the PointTransformerConv (left) and EdgeConv (right) employing the 84 automatically detected landmarks for the gender classification and age prediction tasks, respectively. As can be observed, the approach is capable of learning a representation of the 3D craniofacial morphology of the subjects that effectively stratifies different morphological patterns.

Based on the findings from the experiments conducted on the FDD dataset (cfr. Tables 1, 2), we limit this second phase of evaluation to the two best-performing models: PointTransformerConv and DynamicEdgeConv. Additionally, given the non-critical impact of the connectivity parameter k (cfr. Fig. 2), we set it to 32 as a balanced choice. Experiments were performed using 5-fold cross-validation. Furthermore, on the Facescape dataset, two additional experimental settings were implemented using a leave-one-group-out strategy. In the first scenario (Leave One Emotion Out), we excluded all samples from the same emotion group in each iteration of the cross-validation. This setting provides insights into the models' ability to generalize to variations in facial expressions. In the second scenario (Leave Subject Group Out), we excluded a subset of subjects in each iteration and tested on unseen identities.

The results on these datasets are presented in Table 3. As it can be observed, the accuracy on the DAD-3DH dataset is generally significantly lower compared to the other datasets. This is not surprising, as the face data in DAD-3DH was not natively acquired through 3D scanning; instead, depth information is automatically inferred from 2D images, thus approximating the 3D face geometry.

The overall results (see Tables 1, 2, 3) reveal a significant discrepancy between semantic landmarks and those obtained through a purely geometry-based, non-semantic sampling method. Across all models and configurations, the semantic landmark sets (50 or 84) consistently outperform the sampled set in both tasks. This finding highlights a key insight: facial keypoint selection should be guided by semantic principles rather than purely geometric criteria to achieve a more meaningful characterization of facial morphology, incorporating high-level information beyond spatial distribution alone.

Notably, while the choice of a semantic keypoint set plays a role, its specific composition appears to be less critical. Although manually placed anthropometric landmarks are considered the gold standard in facial anthropometry, the 84 automatically extracted landmarks achieve performance comparable to the 50 anthropometric landmarks across the evaluated tasks. This suggests that automatically extracted keypoints may serve as a faster, more scalable, and cost-effective alternative to manual landmark selection, without compromising representational power.

Qualitatively, the effectiveness of the adopted approaches is assessed by examining the learned model representations as points on a high-dimensional manifold. To visualize this manifold, we use UMAP [42]. Fig. 3 presents 3D UMAP projections of the best EdgeConv and PointTransformerConv models for gender classification and age prediction tasks.

Table 1
Gender Classification on the FDD Dataset: F1 scores obtained by varying the set of landmarks for GNNs node definition (first column), the graph connectivity k (first row), and the model (second column). The best result for each set of landmarks across different models is underlined. The value in bold indicates the overall best result.

	Model/k	3	4	8	16	32	50	84	128
50	GCNN	81.65 ± 2.08	82.66 ± 3.17	82.36 ± 1.40	82.11 ± 1.43	81.69 ± 0.70	81.82 ± 0.73	–	–
	GAT	79.75 ± 3.12	78.02 ± 2.67	78.73 ± 2.12	80.54 ± 4.21	82.79 ± 4.36	83.56 ± 1.77	–	–
	TConv	81.69 ± 0.81	80.52 ± 2.55	82.43 ± 2.64	82.33 ± 2.97	81.38 ± 1.48	79.53 ± 1.92	–	–
	EdgeConv	84.42 ± 2.89	83.99 ± 3.21	84.87 ± 2.44	84.07 ± 4.08	86.01 ± 1.42	83.11 ± 3.92	–	–
	PointTConv	90.40 ± 2.78	91.78 ± 3.85	91.90 ± 5.96	93.63 ± 2.79	93.27 ± 2.07	92.56 ± 3.98	–	–
84	GCNN	93.03 ± 2.35	91.71 ± 3.88	92.33 ± 4.16	90.52 ± 4.42	92.43 ± 2.63	90.34 ± 4.08	87.92 ± 3.96	–
	GAT	83.72 ± 2.65	81.29 ± 2.89	84.53 ± 3.12	86.38 ± 2.81	89.28 ± 3.55	83.24 ± 4.57	81.39 ± 0.82	–
	TConv	86.16 ± 3.21	89.88 ± 3.20	89.49 ± 1.18	86.60 ± 3.21	87.28 ± 3.53	87.88 ± 2.76	87.62 ± 4.59	–
	EdgeConv	89.37 ± 5.42	89.44 ± 4.85	90.56 ± 3.82	93.05 ± 4.46	94.28 ± 4.01	92.61 ± 5.14	**94.33 ± 1.01**	–
	PointTConv	94.54 ± 5.58	92.91 ± 4.58	91.47 ± 2.80	90.15 ± 7.42	93.60 ± 4.42	92.72 ± 2.23	90.42 ± 3.75	–
128	GCNN	81.79 ± 1.06	81.11 ± 0.66	80.68 ± 0.83	81.56 ± 1.45	81.80 ± 0.42	80.00 ± 1.72	78.29 ± 5.52	77.88 ± 6.14
	GAT	79.52 ± 1.65	77.21 ± 0.56	72.66 ± 1.32	75.43 ± 3.48	65.67 ± 2.44	75.49 ± 3.95	77.23 ± 4.53	75.22 ± 2.49
	TConv	80.81 ± 1.00	81.41 ± 0.63	81.78 ± 0.45	79.19 ± 2.89	79.41 ± 1.08	77.05 ± 4.22	80.69 ± 0.46	81.41 ± 0.63
	EdgeConv	82.76 ± 1.34	81.76 ± 0.61	80.18 ± 1.97	81.53 ± 4.16	81.57 ± 2.65	79.76 ± 4.77	79.72 ± 2.79	79.65 ± 2.47
	PointTConv	82.81 ± 4.16	81.05 ± 3.72	79.02 ± 2.91	82.96 ± 4.71	80.30 ± 1.67	77.89 ± 4.53	83.76 ± 3.20	82.56 ± 2.59

Table 2
Age regression on the FDD Dataset: Mean Absolute Error (MAE) obtained by varying the set of landmarks for GNNs node definition (first column), connectivity k (first row), and model (second column). The best result for each set of landmarks across different models is underlined. The value in bold indicates the overall best result.

	Model/k	3	4	8	16	32	50	84	128
50	GCNN	7.34 ± 0.95	6.87 ± 0.76	6.78 ± 0.67	6.86 ± 0.64	6.99 ± 0.58	7.35 ± 0.58	–	–
	GAT	8.17 ± 1.11	8.25 ± 1.03	7.78 ± 0.89	7.11 ± 0.45	6.89 ± 0.76	6.24 ± 0.49	–	–
	TConv	9.56 ± 1.23	8.53 ± 1.34	7.87 ± 1.01	7.98 ± 0.63	8.11 ± 0.78	7.87 ± 0.52	–	–
	EdgeConv	6.21 ± 0.51	5.96 ± 0.68	5.99 ± 0.68	5.77 ± 0.49	5.72 ± 0.90	6.07 ± 0.97	–	–
	PointTConv	5.20 ± 0.38	5.37 ± 0.53	5.57 ± 0.76	5.27 ± 0.62	4.77 ± 0.70	4.88 ± 0.69	–	–
84	GCNN	5.23 ± 0.18	5.27 ± 0.35	5.63 ± 0.17	5.01 ± 0.85	4.95 ± 0.50	5.13 ± 0.42	5.34 ± 0.39	–
	GAT	8.04 ± 0.56	8.25 ± 0.45	6.54 ± 1.42	6.11 ± 1.09	5.87 ± 1.25	5.56 ± 1.99	5.78 ± 0.35	–
	TConv	9.91 ± 2.13	10.33 ± 1.73	12.23 ± 1.28	6.25 ± 2.11	8.23 ± 1.38	6.29 ± 1.83	5.12 ± 0.78	–
	EdgeConv	4.29 ± 0.36	4.65 ± 0.51	4.24 ± 0.38	3.92 ± 0.36	4.12 ± 0.42	**3.95 ± 0.25**	4.12 ± 0.11	–
	PointTConv	4.15 ± 0.31	4.05 ± 0.40	4.24 ± 0.36	4.40 ± 0.43	4.31 ± 0.38	4.25 ± 0.53	4.15 ± 0.66	–
128	GCNN	9.65 ± 1.09	9.47 ± 1.09	8.96 ± 0.94	9.27 ± 1.19	12.40 ± 2.86	9.92 ± 2.66	9.40 ± 0.75	9.16 ± 1.02
	GAT	10.11 ± 3.22	9.87 ± 1.67	8.89 ± 1.29	7.43 ± 1.13	9.34 ± 1.81	9.11 ± 2.15	8.45 ± 1.45	9.61 ± 1.89
	TConv	17.11 ± 2.45	13.42 ± 3.12	15.83 ± 2.55	15.45 ± 2.75	12.76 ± 2.49	13.15 ± 1.29	11.23 ± 1.65	9.56 ± 1.39
	EdgeConv	8.25 ± 0.84	7.76 ± 1.27	8.59 ± 0.88	8.41 ± 0.64	8.03 ± 1.05	7.93 ± 0.96	8.09 ± 1.26	8.30 ± 1.01
	PointTConv	8.44 ± 0.96	8.00 ± 0.45	8.20 ± 1.14	7.94 ± 0.70	7.85 ± 0.55	7.87 ± 0.54	7.97 ± 0.29	7.82 ± 0.14

Table 3
Gender classification on the Facescape and DAD-3DH Datasets: F1 scores obtained by varying the set of landmarks for GNNs node definition (first column), and the model (second column). For each set of landmarks, the best result across different models is underlined, while the overall best result for each experimental setting is highlighted in bold.

		Facescape			DAD3D-Heads
		Traditional K-Fold	Leave Subject Group Out	Leave One Emotion out	Traditional K-Fold
50	EdgeConv	96.72 ± 0.54	91.61 ± 0.50	97.50 ± 0.74	68.25 ± 1.84
	PointTConv	96.16 ± 1.11	93.39 ± 0.53	96.30 ± 0.69	68.48 ± 3.35
84	EdgeConv	**96.92 ± 0.47**	91.34 ± 1.55	**97.10 ± 0.95**	64.49 ± 6.94
	PointTConv	96.10 ± 0.12	**96.17 ± 0.10**	96.17 ± 1.04	**74.23 ± 2.17**
128	EdgeConv	84.01 ± 1.23	76.95 ± 2.28	81.50 ± 3.02	45.97 ± 20.71
	PointTConv	85.98 ± 0.60	81.89 ± 1.88	85.08 ± 0.96	61.27 ± 8.48

3.4. Significance testing

The results summarized in the previous section are substantiated here using appropriate statistical assessment techniques. This study verifies whether the differences in a given metric are statistically significant or occur by chance. Typically, this involves Null Hypothesis Statistical Testing (NHST), commonly employed for rigorous performance evaluation of classification algorithms [43]. Recently, Benavoli et al. [44] proposed using Bayesian estimation techniques to assess the significance of performance differences between machine learning algorithms across datasets.

Following a similar approach, this work evaluates the performance differences yielded by various sets of landmarks used for facial morphological representation learning through a Bayesian non-parametric method that extends the Wilcoxon signed-rank test [45]. Unlike frequentist NHST, Bayesian estimation examines the actual probability of one method outperforming or equaling another. This comparison is conducted using the Region of Practical Equivalence (ROPE), an interval within which performance differences between approaches are deemed negligible for a given metric. As suggested by Kruschke and Liddell [46], ROPE is typically set as half the value of a negligible magnitude for the metric.

For the MAE metric, which measures the average error in age prediction, performance differences below 1 year are considered negligible; thus, we set $rope = 0.5$. For the F1 metric, which measures gender classification accuracy, $rope = 0.05$. Fig. 4 illustrates the results, showing posterior distribution samples of the Bayesian signed-rank test on the simplex for the F1 score (top right) and MAE (bottom right). Each vertex of the triangle represents a scenario where an approach is more likely to produce higher metric values, equivalent results, or lower values relative to another.

The findings are summarized as follows: the distribution of gender classification accuracies (F1 scores) for models using either the 50 manually selected anthropometric landmarks or the 84 automatically

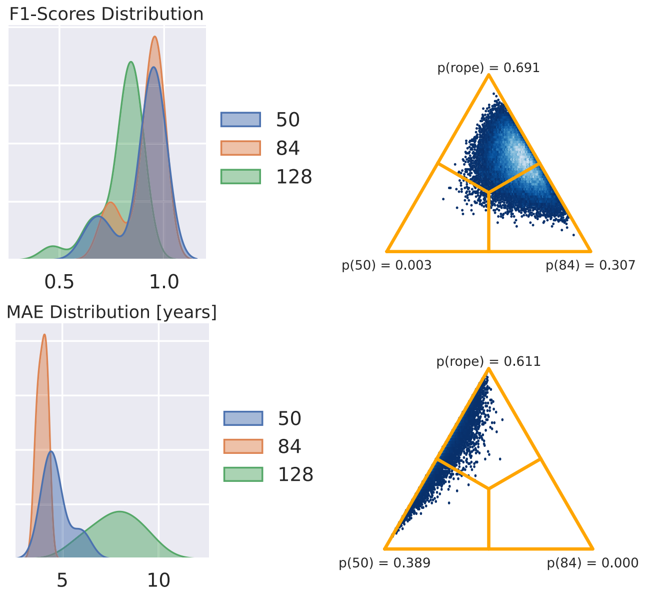

Fig. 4. Statistical significance testing. Probability estimates for a given set of facial landmarks to yield better performance on the evaluated tasks: gender classification (top row) and age estimation (bottom row). The first column displays the empirical distributions of the obtained F1 scores (for gender classification) and MAE values (for age estimation). The second column shows the posterior samples for the Bayesian Sign-Rank Test on the simplex, illustrating the probability that one set of landmarks produces higher (or equivalent) metric values compared to another.

detected landmarks indicates similar facial morphological representation capabilities (see Fig. 4, top left). The Bayesian signed-rank test (Fig. 4, top right) confirms this, with results falling within the ROPE with a probability of 0.691. Notably, the test also reveals a non-negligible probability (0.307) that the automatically selected landmarks yield better performance. Similar conclusions can be drawn for the age prediction results (Fig. 4, bottom row).

4. Conclusions

This study explores the use of GNNs for analyzing 3D facial morphology, adopting facial landmarks as graph nodes to represent the intrinsic structure of 3D face scans. The research compares three types of facial keypoints: a well established set of 50 anthropometric 3D facial landmarks selected by expert anthropometrists via visual and tactile assessment, a set of 84 automatically detected 3D facial landmarks from state-of-the-art models, and a broader set of 128 facial keypoints obtained by sampling the original point cloud via Farthest Point Sampling (FPS). The primary goals are to evaluate the use of semantically meaningful landmarks versus geometry-based sampling (FPS) and determine whether automated landmarks can serve as a viable alternative to anthropometric landmarks. Gender classification and age regression tasks are used as benchmarks to characterize facial morphology. Among the GNN architectures examined, PointTransformerConv and DynamicEdgeConv consistently demonstrated superior performance across all tasks and datasets. The key findings can be summarized as follows.

1. Semantic Landmarks vs. Geometry-Based Sampling (RQ1). The study reveals the clear advantages of semantically meaningful landmarks, such as anthropometric or automatically detected keypoints, over geometry-based sampled landmarks. Across all models and tasks, semantically meaningful landmarks consistently achieved better performance.
2. Automatic vs. Anthropometric Landmarks (RQ2). The results demonstrate that automatically extracted landmarks perform comparably to, and sometimes better than, manually annotated anthropometric landmarks. Statistical testing confirms this performance parity, suggesting that automated landmark detection offers a scalable, cost-effective alternative to manual annotation. Automated techniques accelerate the data processing pipeline, making them particularly suitable for large-scale studies where manual annotation is impractical. Crucially, when paired with advanced GNN architectures, automated landmarks maintain the representational power of facial morphology data, paving the way for broader adoption in research and clinical settings.

In terms of implications and future directions, these results have significant implications for the assessment of facial morphology in 3D. While manual anthropometric landmark selection remains vital in sensitive applications, such as forensic investigations and specific clinical diagnostics, the study demonstrates the potential for automated, scalable solutions in broader contexts. Applications of these methodologies extend beyond gender classification and age regression to include, facial expression recognition, physical anthropology research, cognitive science studies and clinical applications in plastic surgery, orthodontics, and syndrome diagnosis. Overall, this research establishes a foundation for more efficient approaches to 3D facial analysis, reducing reliance on manual annotation while ensuring high performance. Future work could explore additional facial morphology tasks, further validating the versatility and scalability of automated landmarks in conjunction with advanced GNN architectures.

CRediT authorship contribution statement

Giuseppe Maurizio Facchi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation. **Giuliano Grossi:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Alessandro D'Amelio:** Visualization, Resources, Methodology, Data curation, Conceptualization. **Francesco Agnelli:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Chiarella Sforza:** Writing – review & editing, Validation, Supervision, Project administration, Formal analysis. **Gianluca Martino Tartaglia:** Writing – original draft, Validation, Supervision, Project administration, Conceptualization. **Raffaella Lanzarotti:** Writing – review & editing, Writing – original draft, Software, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patrec.2025.04.028.

Data availability

The authors do not have permission to share data.

References

[1] F. Becattini, L. Berlincioni, L. Cultrera, A. Del Bimbo, Neuromorphic face analysis: A survey, Pattern Recognit. Lett. 187 (2025) 42–48.
[2] C. Lv, Z. Wu, X. Wang, Z. Dan, M. Zhou, Ethnicity classification by the 3d discrete landmarks model measure in Kendall shape space, Pattern Recognit. Lett. 129 (2020) 26–32.
[3] G. Boccignone, A. D'Amelio, O. Ghezzi, G. Grossi, R. Lanzarotti, An evaluation of non-contact photoplethysmography-based methods for remote respiratory rate estimation, Sensors 23 (2023) 3387.

[4] S. Patania, G. Boccignone, S. Buršić, A. D'Amelio, R. Lanzarotti, Deep graph neural network for video-based facial pain expression assessment, in: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, 2022, pp. 585–591.

[5] V. Pucciarelli, D. Gibelli, C. Mastella, S. Bertoli, K. Alberti, M. De Amicis, C. Dolci, A. Battezzati, G. Baranello, et al., 3D facial morphology in children affected by spinal muscular atrophy type 2 (smaii), Eur. J. Orthod. 42 (2020) 500–508.

[6] R.M. Winter, What's in a face? Nature Genet. 12 (1996) 124–129.

[7] J. Thevenot, M.B. López, A. Hadid, A survey on computer vision for assistive medical diagnosis from faces, IEEE J. Biomed. Heal. Inform. 22 (2017) 1497–1511.

[8] N.H. Vu, N.M. Trieu, H.N. Anh Tuan, T.D. Khoa, N.T. Thinh, Facial anthropometric, landmark extraction, and nasal reconstruction technology, Appl. Sci. 12 (2022) 9548.

[9] D. Gibelli, C. Dolci, A. Cappella, C. Sforza, Reliability of optical devices for three-dimensional facial anatomy description: A systematic review and meta-analysis, Int. J. Oral Maxillofac. Surg. 49 (2020) 1092–1106.

[10] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S.Y. Philip, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (2020) 4–24.

[11] C. Chen, Y. Wu, Q. Dai, H.-Y. Zhou, M. Xu, S. Yang, X. Han, Y. Yu, A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[12] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Adv. Neural Inf. Process. Syst. 30 (2017).

[13] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017, URL: https://openreview.net/forum?id=SJU4ayYgl.

[14] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, Y. Sun, Masked label prediction: Unified message passing model for semi-supervised classification, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, 2021.

[15] Z. Zhang, P. Cui, W. Zhu, Deep learning on graphs: A survey, IEEE Trans. Knowl. Data Eng. 34 (2020) 249–270.

[16] S. Jiang, Q. Feng, H. Li, Z. Deng, Q. Jiang, Attention based multi-task interpretable graph convolutional network for Alzheimer's disease analysis, Pattern Recognit. Lett. 180 (2024) 1–8.

[17] H. Huang, L. Wang, X. Li, S. Yuan, C. Wen, Y. Hao, Y. Fang, Learning to learn point signature for 3d shape geometry, Pattern Recognit. Lett. (2024).

[18] S. Liu, S. Huang, W. Fu, J.C.-W. Lin, A descriptive human visual cognitive strategy using graph neural network for facial expression recognition, Int. J. Mach. Learn. Cybern. 15 (2024) 19–35.

[19] R.R. Paulsen, K.A. Juhl, T.M. Haspang, T. Hansen, M. Ganz, G. Einarsson, Multi-view consensus cnn for 3d facial landmark placement, in: Asian Conference on Computer Vision, Springer, 2018, pp. 706–719.

[20] Y. Liu, C.H. Kau, L. Talbert, F. Pan, Three-dimensional analysis of facial morphology, J. Craniofac. Surg. 25 (2014) 1890–1894.

[21] M.J. Kesterke, Z.D. Raffensperger, C.L. Heike, M.L. Cunningham, J.T. Hecht, C.H. Kau, N.L. Nidey, L.M. Moreno, G.L. Wehby, M.L. Marazita, et al., Using the 3d facial norms database to investigate craniofacial sexual dimorphism in healthy children, adolescents, and adults, Biol. Sex Differ. 7 (2016) 1–14.

[22] A.K. Ingale, A.A. Leema, H. Kim, J.D. Udayan, Automatic 3d facial landmark-based deformation transfer on facial variants for blendshape generation, Arab. J. Sci. Eng. 48 (2023) 10109–10123.

[23] S. Liang, J. Wu, S.M. Weinberg, L.G. Shapiro, Improved detection of landmarks on 3d human face data, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, IEEE, 2013, pp. 6482–6485.

[24] L.A. Jeni, J.F. Cohn, T. Kanade, Dense 3d face alignment from 2d videos in real-time, in: 2015 11th IEEE intìl Conf. on Automatic Face and Gesture Recognition, Vol. 1, FG, IEEE, 2015, pp. 1–8.

[25] S. Tulyakov, N. Sebe, Regressing a 3d face shape from a single image, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3748–3755.

[26] Y. Wang, M. Cao, Z. Fan, S. Peng, Learning to detect 3d facial landmarks via heatmap regression with graph convolutional network, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 2595–2603.

[27] J. Burger, G. Blandano, G.M. Facchi, R. Lanzarotti, 2S-sgcn: A two-stage stratified graph convolutional network model for facial landmark detection on 3d data, Comput. Vis. Image Underst. 250 (2025) 104227.

[28] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, J. Lu, Point-bert: Pre-training 3d point cloud transformers with masked point modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19313–19322.

[29] H. Fan, H. Su, L. J. Guibas, A point set generation network for 3d object reconstruction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 605–613.

[30] Z. Lyu, J. Wang, Y. An, Y. Zhang, D. Lin, B. Dai, Controllable mesh generation through sparse latent point diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 271–280.

[31] K. Crane, C. Weischedel, M. Wardetzky, The heat method for distance computation, Commun. ACM 60 (2017) 90–99.

[32] A. Loukas, What graph neural networks cannot learn: depth vs width, in: International Conference on Learning Representations, 2020.

[33] R.B. Rusu, N. Blodow, M. Beetz, Fast point feature histograms (fpfh) for 3d registration, in: 2009 IEEE International Conference on Robotics and Automation, IEEE, 2009, pp. 3212–3217.

[34] S. Yun, M. Jeong, R. Kim, J. Kang, H.J. Kim, Graph transformer networks, Adv. Neural Inf. Process. Syst. 32 (2019).

[35] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al., Graph attention networks, Stat 1050 (2017) 10–48550.

[36] Y. Wang, Y. Sun, Z. Liu, S.E. Sarma, M.M. Bronstein, J.M. Solomon, Dynamic graph cnn for learning on point clouds, ACM Trans. Graph. (Tog) 38 (2019) 1–12.

[37] H. Zhao, L. Jiang, J. Jia, P.H. Torr, V. Koltun, Point transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16259–16268.

[38] C.R. Qi, H. Su, K. Mo, L. J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.

[39] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, X. Cao, Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 601–610.

[40] T. Martyniuk, O. Kupyn, Y. Kurlyak, I. Krashenyi, J. Matas, V. Sharmanska, Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3d head alignment from a single image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20942–20952.

[41] T. Li, T. Bolkart, M.J. Black, H. Li, J. Romero, Learning a model of facial shape and expression from 4D scans, ACM Trans. Graph. 36 (2017) http://dx.doi.org/10.1145/3130800.3130813, (Proc. SIGGRAPH Asia).

[42] L. McInnes, J. Healy, N. Saul, L. Großberger, Umap: Uniform manifold approximation and projection, J. Open Source Softw. 3 (2018) 861.

[43] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[44] A. Benavoli, G. Corani, J. Demšar, M. Zaffalon, Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis, J. Mach. Learn. Res. 18 (2017) 2653–2688.

[45] A. Benavoli, G. Corani, F. Mangili, M. Zaffalon, F. Ruggeri, A Bayesian wilcoxon signed-rank test based on the Dirichlet process, in: International Conference on Machine Learning, PMLR, 2014, pp. 1026–1034.

[46] J.K. Kruschke, T.M. Liddell, The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective, Psychon. Bull. Rev. 25 (2018) 178–206.