Learning Italian Hand Gesture Culture Through an Automatic Gesture Recognition Approach

future internet

MDPI

Article

Learning Italian Hand Gesture Culture Through an Automatic Gesture Recognition Approach

Chiara Innocente [1], Giorgio Di Pisa [2], Irene Lionetti [2], Andrea Mamoli [2], Manuela Vitulano [2], Giorgia Marullo [1], Simone Maffei [1], Enrico Vezzetti [1] and Luca Ulrich [1,*]

[1] Management and Production Engineering, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129 Torino, Italy; chiara.innocente@polito.it (C.I.); giorgia.marullo@polito.it (G.M.); enrico.vezzetti@polito.it (E.V.)

[2] Biomedical Engineering, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129 Torino, Italy; s329317@studenti.polito.it (I.L.)

* Correspondence: luca.ulrich@polito.com

Abstract

Italian hand gestures constitute a distinctive and widely recognized form of nonverbal communication, deeply embedded in everyday interaction and cultural identity. Despite their prominence, these gestures are rarely formalized or systematically taught, posing challenges for foreign speakers and visitors seeking to interpret their meaning and pragmatic use. Moreover, their ephemeral and embodied nature complicates traditional preservation and transmission approaches, positioning them within the broader domain of intangible cultural heritage. This paper introduces a machine learning–based framework for recognizing iconic Italian hand gestures, designed to support cultural learning and engagement among foreign speakers and visitors. The approach combines RGB–D sensing with depth-enhanced geometric feature extraction, employing interpretable classification models trained on a purpose-built dataset. The recognition system is integrated into a non-immersive virtual reality application simulating an interactive digital totem conceived for public arrival spaces, providing tutorial content, real-time gesture recognition, and immediate feedback within a playful and accessible learning environment. Three supervised machine learning pipelines were evaluated, and Random Forest achieved the best overall performance. Its integration with an Isolation Forest module was further considered for deployment, achieving a macro-averaged accuracy and F1-score of 0.82 under a 5-fold cross-validation protocol. An experimental user study was conducted with 25 subjects to evaluate the proposed interactive system in terms of usability, user engagement, and learning effectiveness, obtaining favorable results and demonstrating its potential as a practical tool for cultural education and intercultural communication.

Keywords: intangible cultural heritage; Italian culture; gesture recognition; human–computer interaction; machine learning; virtual reality

Academic Editor: Diego Vergara

Received: 27 February 2026
Revised: 20 March 2026
Accepted: 21 March 2026
Published: 24 March 2026

1. Introduction

Italy's rich tradition of expressive hand gestures plays a central role in everyday communication and represents a distinctive element of national cultural identity [1,2]. Beyond their auxiliary function to spoken language, Italian gestures often convey complete semantic meanings on their own, encoding emotions, intentions, and social conventions that are deeply rooted in historical and regional contexts [3]. Research on symbolic gestures demonstrates that gestures can be analyzed as culturally codified signals with semantic

structure, forming an organized gesturality rather than random hand movements. Seminal work by Telmon [4] systematically categorized Italian gestures into an analogical and alphabetical gestuario, highlighting the cultural richness, symbolic value, and communicative complexity of this nonverbal language [4]. Moreover, empirical work on Southern Italian conversation has shown that many gestures operate as illocutionary and discourse structure markers that contribute meaning independently of speech content [5]. Cross-cultural research comparing Italian and Swedish speakers confirms that Italians gesture more frequently and use gestures differently in narrative contexts, reinforcing the idea that gesture use is shaped by cultural communication norms [6]. Moreover, studies in language and communication contexts emphasize the pedagogical relevance of gestures in Italian communicative situations [7].

Due to their pervasiveness and recognizability, Italian hand gestures have become internationally known stereotypes of Italian expressiveness. The enduring symbolic value of Italian gestural communication is also evident in contemporary cultural representation. For example, the opening ceremony of the Milano–Cortina 2026 Winter Olympics, whose artistic program aimed to showcase Italian cultural identity, included a segment presenting iconic hand gestures as a recognizable form of everyday communication, highlighting their status as markers of national expressiveness. At the same time, this widespread visibility has raised growing interest in their recognition as a form of intangible cultural heritage, in line with UNESCO's definition of cultural practices, expressions, and knowledge that communities recognize as part of their cultural legacy.

Preserving and transmitting these nonverbal forms of communication is particularly challenging, as gestures are inherently ephemeral, context-dependent, and traditionally learned through direct social exposure rather than formal instruction [8]. This challenge is especially evident for tourists and non-native speakers, who often struggle to decode the meaning of Italian gestures and consequently miss important pragmatic nuances of everyday interactions. As a result, opportunities for deeper cultural immersion and mutual understanding are frequently lost. In this context, educational tools capable of supporting the learning and interpretation of culturally specific gestures could play a significant role in promoting intercultural communication and enhancing cultural engagement.

Recent advances in digital technologies have opened new opportunities for the preservation and dissemination of cultural heritage, especially for embodied and performative forms of knowledge [9]. Interactive multimedia systems, as well as virtual and augmented reality applications, have enabled a shift from passive consumption toward participatory and immersive experiences, allowing users to actively explore cultural practices [10,11]. Such approaches are particularly suitable for intangible cultural heritage, where gestures and performative expressions can be represented as dynamic processes rather than static artifacts [12,13].

Within this technological landscape, gesture recognition has emerged as an active research field, initially driven by applications in human–computer interaction, sign language interpretation, and assistive technologies.

Gesture recognition has evolved from early probabilistic approaches based on visual data, such as hidden Markov models for sign language interpretation [14,15], toward more robust methods addressing challenges related to detection, temporal segmentation, and real-time performance [16,17]. Subsequent research introduced geometric and kinematic representations of the hand, including joint angles and inter-joint distances, supported by both vision-based and sensor-based systems, as well as wearable solutions combined with classical machine learning techniques [18–20]. The introduction of depth-sensing technologies enabled explicit three-dimensional modeling of human pose and hand articulations, significantly improving robustness and real-time capabilities [21–23]. More recently, hybrid

approaches combining geometric representations with learning-based models have further advanced the field, with frameworks such as MediaPipe and OpenPose enabling real-time hand pose estimation from RGB data [24,25]. These approaches are increasingly integrated with machine learning classifiers to achieve efficient and accurate gesture recognition on consumer devices, including RGB–D solutions leveraging landmark-based features to improve the discrimination of geometrically similar hand configurations [26–29].

Despite significant progress in gesture recognition, most existing systems focus on general-purpose gestures or formalized languages, such as sign languages [30], and are rarely designed with cultural specificity as a primary objective. Only a limited number of studies have addressed the recognition of culture-specific gestures, often in non-European contexts [31,32]. For example, recent work has explored the automatic recognition of traditional dance movements associated with intangible cultural heritage, leveraging multi-feature fusion strategies that integrate skeletal, spatiotemporal, and deep features to classify culturally grounded gestures from diverse dance traditions [33]. Similarly, Kuchipudi classical dance mudras have been recognized in real time using MediaPipe landmarks combined with a Support Vector Machine classifier [34]. These works demonstrate the feasibility of applying artificial intelligence techniques to culturally grounded gestures, but they remain largely disconnected from broader discussions on cultural heritage promotion and learning.

Against this background, there is a growing need for systems that are interpretable, computationally lightweight, and explicitly sensitive to cultural context, in order to effectively support the learning, transmission, and dissemination of intangible cultural heritage.

The aim of this work is to present a gesture recognition system designed to support the learning of iconic Italian hand gestures, explicitly conceived as an interactive tool for cultural education and engagement for foreign speakers and visitors. The contribution of this work lies in the design and integration of a lightweight recognition pipeline within an interactive learning environment specifically tailored to the transmission of culturally embedded nonverbal communication. In particular, the system enables users to actively reproduce gestures and receive immediate recognition feedback, facilitating the understanding of Italian gestural communication that is commonly encountered in everyday interactions but rarely formalized in traditional language learning materials. The proposed approach combines RGB–D sensing with geometric feature extraction and interpretable machine learning classifiers. In contrast to deep learning architectures that typically require large-scale datasets for stable training, the proposed lightweight pipeline is designed to operate effectively in small-data scenarios while maintaining computational efficiency and transparency. The recognition pipeline relies on geometric descriptors derived from hand landmark reconstruction, capturing the spatial configuration of the fingers through inter-joint distances that approximate the anatomical structure of the hand. This representation provides an interpretable description of hand posture and enables efficient classification using lightweight models. The lightweight recognition pipeline is integrated into an interactive virtual environment designed to support playful, immersive, and user-centered learning experiences, with a strong emphasis on accessibility and deployment on consumer hardware. These design choices make the system particularly suitable for deployment in real-world scenarios where explainability, low latency, and limited computational resources are critical, opening up opportunities for practical applications in public and semi-public spaces such as airport terminals, transportation hubs, tourist information centers, museums, and cultural venues, where foreign visitors could be introduced to iconic Italian gestures in an intuitive and engaging manner. In these contexts, the system can function as an educational interface promoting cultural awareness, facilitating intercultural communication, and enhancing the overall visitor experience through embodied and interactive learning.

The remainder of this paper is structured as follows. Section 2 describes the data acquisition process, feature extraction pipeline, machine learning methodology, and their integration into an interactive virtual environment. Section 3 presents the experimental evaluation, Section 4 discusses the results in relation to existing literature and cultural heritage applications, and finally Section 5 draws conclusions.

2. Materials and Methods

The overall methodological workflow of the proposed gesture recognition architecture is summarized in Figure 1, illustrating the main processing stages from RGB–D acquisition and hand landmark detection to feature extraction, machine learning classification, and integration into the interactive Unity-based application.

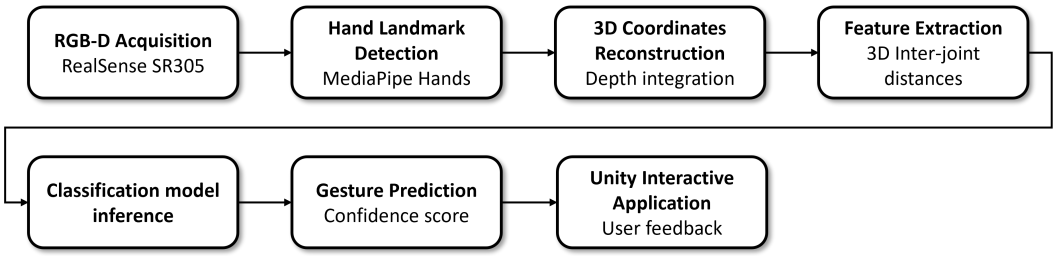


Figure 1. Overview of the proposed gesture recognition pipeline, illustrating the main processing stages from RGB–D acquisition and hand landmark detection to feature extraction, machine learning classification, and integration into the interactive Unity application.

The methodological approach is described as follows. Section 2.1 presents the dataset acquisition procedure, including the acquisition protocol and the applied preprocessing steps. Section 2.2 details the feature extraction process, while Section 2.3 describes the implemented classification pipeline. Finally, Section 2.4 introduces the developed VR interaction environment.

2.1. Dataset Acquisition

The dataset was specifically designed to capture a set of iconic Italian hand gestures that are widely recognized in everyday communication and frequently encountered by foreign speakers, yet often difficult to interpret without cultural familiarity. Four gestures were selected based on their cultural relevance, semantic distinctiveness, and suitability for recognition using RGB–D sensing. The selected gestures, referenced to Telmon [4] work, represent commonly used communicative acts in Italian nonverbal interaction and were chosen to minimize semantic overlap while preserving variability in hand configuration and motion. A visual representation of the selected gestures and a description of their possible meaning in Italian nonverbal communication is provided in Table 1.

The acquisition protocol was designed to ensure consistency across recordings while allowing natural variability in gesture execution. Data were collected from ten participants, each performing the selected gestures three times with the dominant hand and subsequently repeating them with the non-dominant hand, resulting in multiple repetitions per gesture and participant. This acquisition protocol was designed to increase both intra-subject variability and handedness variability within the dataset. In addition to the target gestures, each participant was also asked to perform two unconstrained hand gestures not belonging to the predefined categories. These additional gestures were included to populate an extra class representing non-relevant or unknown gestures (Class *Others*), enabling the model to learn to reject hand configurations that do not correspond to the target cultural gestures. Participants were instructed to execute each gesture in a natural manner, without strict constraints on speed or amplitude, in order to reflect realistic usage scenarios.

Table 1. Selected Italian hand gestures included in the dataset and their cultural meanings.

Gesture Name	Execution and Cultural Meaning
Question Mark Hand	The gesture is performed by bringing the fingertips together with the palm oriented upward, often accompanied by a slight vertical motion. It is used to express questioning, disbelief, or impatience (e.g., "What are you saying?", "What do you want?", or "What does this mean?").
Horns	The gesture is executed by extending the index and little fingers while keeping the remaining fingers folded toward the palm. It is commonly employed as a protective sign against bad luck or the evil eye, particularly when oriented downward or executed discreetly.
Crossed Fingers	The gesture consists of crossing the middle finger over the index finger, with the remaining fingers folded. It is commonly associated with wishing good luck, expressing hope, or warding off misfortune.
Perfection	The gesture is formed by creating a circular shape between the thumb and index finger while extending the other fingers. In Italian communication, particularly in southern regions, it emphasizes correctness and precision, differing from the "OK" gesture by conveying instruction rather than approval and potentially leading to cross-cultural misinterpretation.

All recordings, lasting three seconds each, were conducted indoors under uniform ambient illumination and against a neutral background, with the participant's torso visible behind the hand. Participants were seated at an approximate distance of 70 cm from an Intel RealSense SR305 (Intel Corporation, Santa Clara, CA, USA) RGB–D camera, capturing synchronized color and depth streams at a spatial resolution of 640 × 480 pixels and a frame rate of 30 fps. During acquisition, the hand was positioned at a distance between 30 cm and 40 cm from the camera to ensure reliable depth sensing and accurate hand tracking. To reduce potential biases, the order of gesture execution was randomized across participants. The dataset acquisition protocol was reviewed to ensure compliance with ethical standards, and all participants provided informed consent prior to data collection. For each gesture repetition, a single frame was manually selected as the most visually representative of the gesture, typically corresponding to the frame of the most distinctive hand configuration. This choice was motivated by the focus on static hand configurations, which are sufficient to characterize the selected gestures while reducing computational complexity and dataset dimensionality.

Color frames were processed using Python (version 3.9.13) and the MediaPipe Hands model through the MediaPipe Python API (version 0.10.32). The model was configured to detect at most one hand per frame and to output 21 two-dimensional pixel coordinates (u, v) corresponding to the detected hand landmarks. Frames in which no hand was successfully detected were excluded from further analysis. While MediaPipe provides an estimated depth value for each detected landmark, this value is inferred by the model rather than obtained from a direct depth measurement [24]. In this work, the estimated landmark depth was therefore replaced with the corresponding value extracted from the depth map acquired by the RealSense sensor. This integration allowed the reconstruction of accurate three-dimensional landmark coordinates and enabled the computation of true 3D inter-joint distances, providing a more reliable geometric representation of hand posture compared to RGB-only landmark estimation [35]. The use of depth information, therefore, improves the consistency of the geometric representation across variations in

viewpoint, hand orientation, and lighting conditions, providing a more reliable description of hand posture compared to RGB-only landmark estimation. Each raw depth map was subsequently preprocessed to address missing or invalid values. A binary mask of valid depth pixels (depth values > 0) was first generated, and missing values were filled by replacing each zero-valued pixel with the arithmetic mean of its eight orthogonal and diagonal neighbors. The 2D pixel coordinates and their corresponding depth values were then deprojected into camera-centered three-dimensional points $P = (X, Y, Z)$ expressed in meters. Depth values were first converted using the camera depth scale factor of 0.000125 m per unit. The deprojection was performed using a pinhole camera model with radial and tangential distortion correction, defined as:

$$x = \frac{u - c_x}{f_x}, \quad y = \frac{v - c_y}{f_y}, \quad r^2 = x^2 + y^2 \tag{1}$$

$$f(r) = 1 + k_1 r^2 + k_2 r^4 + k_3 r^6 \tag{2}$$

$$u' = x f(r) + 2 p_1 x y + p_2 (r^2 + 2 x^2), \quad y' = y f(r) + 2 p_2 x y + p_1 (r^2 + 2 y^2) \tag{3}$$

$$X = d u', \quad Y = d y', \quad Z = d \tag{4}$$

where $(c_x, c_y) = (309.568, 245.154)$ and $(f_x, f_y) = (474.346, 474.346)$ denote the camera principal point and focal lengths, respectively; $(k_1, k_2, p_1, p_2, k_3) = (0.139663, 0.0914142, 0.00468509, 0.00220023, 0.0654529)$ are the radial and tangential distortion coefficients; and d represents the scaled depth value. All intrinsic parameters were obtained from the camera manufacturer's specifications and were not subject to additional calibration.

At the end of the preprocessing stage, the final dataset consisted of a total of 265 samples, distributed across gesture classes as reported in Table 2. This total reflects both the inclusion of the additional *Others* gesture class and the exclusion of frames in which hand detection was unsuccessful.

Table 2. Composition of the final dataset with number of samples per class.

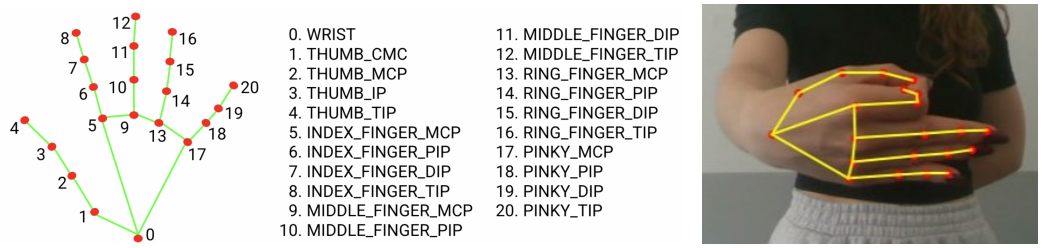
Question Mark Hand	Horns	Crossed Fingers	Perfection	Others
50	55	50	57	53

Each sample comprises the gesture class label, the two-dimensional MediaPipe hand landmark coordinates, and the corresponding depth values extracted from the depth map, which together define the three-dimensional hand configuration used for subsequent feature extraction and classification. Given the relatively small size of the dataset, no further balancing procedures were applied to enforce an equal number of samples across classes.

2.2. Feature Extraction

Following the approach adopted in the MediaPipe Hands framework [24], a subset of twenty-one pairwise distances between hand landmarks was selected as the feature representation. The selected distances follow the anatomical structure of the hand, approximating the skeletal relationships between joints and phalanges. This design choice ensures that the feature representation reflects the physical configuration of the hand posture rather than arbitrary spatial relationships between landmarks, while remaining invariant to global hand orientation and translation (Figure 2). For each sample, three-dimensional Euclidean distances were computed between the selected landmark pairs using their reconstructed 3D coordinates. Distances were expressed in centimeters to improve numerical interpretability. To mitigate the effect of spurious values arising from occasional landmark misplacement or

missing depth measurements, all distances were clipped to a maximum value of 15 cm. The threshold was chosen to remain within the physiologically plausible range of inter-finger distances observed during data collection, while preventing unrealistically large values from influencing the classifier, therefore acting as a robustness mechanism that limits the impact of outliers while preserving valid hand configurations.



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

Figure 2. (**a**) MediaPipe stock image showing the 21 2D hand landmarks with anatomical names and indices; (**b**) RGB frame from the dataset overlaid with the MediaPipe computed landmarks.

The resulting feature vectors were assembled into a single dataset and subsequently standardized feature-wise to zero mean and unit variance. Specifically, each distance feature was transformed by subtracting its empirical mean and dividing by its standard deviation computed over the training data. This normalization step was applied to ensure that all features contributed equally to the learning process, preventing distances with larger numerical ranges from disproportionately influencing the classifier.

As an initial exploratory analysis, a correlation matrix was computed to assess the degree of linear dependence among the selected 3D landmark distance features. This analysis aimed to identify potential redundancy or multicollinearity that could negatively affect classification performance. The results, reported in Figure 3, indicate that the majority of the selected distances exhibit low pairwise correlation, with most coefficients remaining below 0.3 in absolute value. As expected, moderate correlation values can be observed between `WRIST_INDEX_MCP` and `WRIST_THUMB_CMC`, due to the shared reference point at the wrist and the intrinsic kinematic coupling of neighboring finger structures.

As no feature pairs display strong correlation with values above 0.7, all calculated distances were retained without applying dimensionality reduction, as the features do not introduce significant redundancy that could negatively impact model stability or interpretability.

A 5-fold cross-validation strategy was adopted for dataset partitioning and model evaluation. The dataset was randomly divided into five stratified folds, ensuring that each fold preserved the class distribution. At each iteration, one fold was used as the test set, while the remaining four folds were used for training. This process was repeated five times, allowing each sample to be used once for testing and four times for training. A fixed random seed of 42 was used to ensure reproducibility.

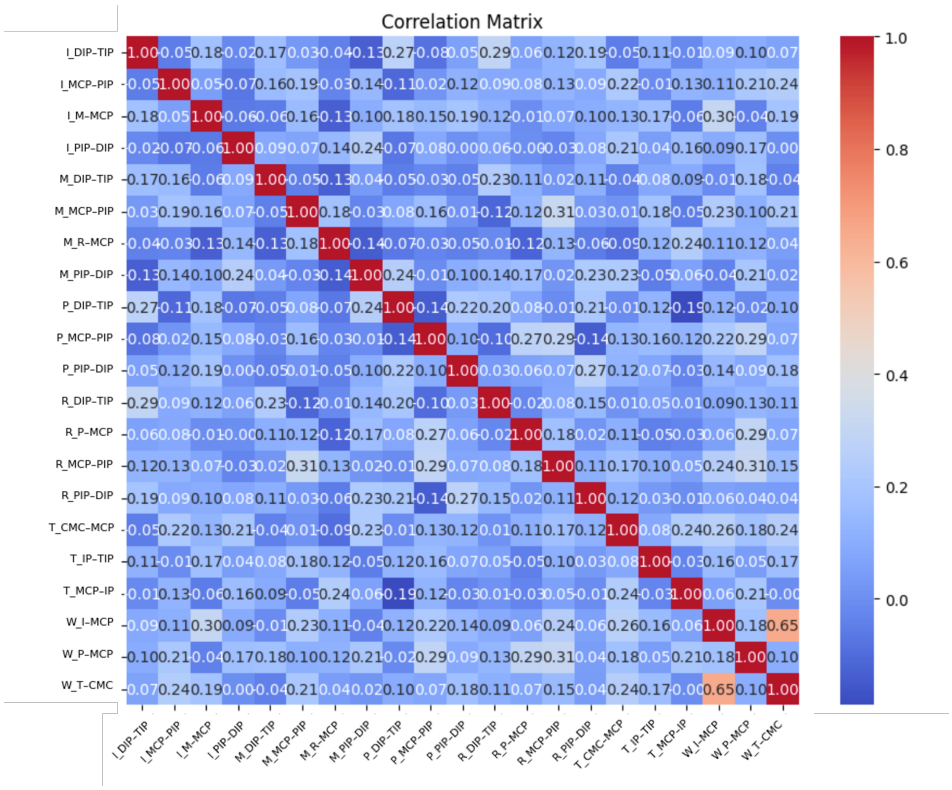


Figure 3. Correlation matrix of the selected 3D hand landmark distance features. Feature labels are abbreviated for readability (I = Index, M = Middle, R = Ring, P = Pinky, T = Thumb, W = Wrist).

2.3. Classification Pipeline

The classification pipeline was designed to ensure a fair, reproducible, and interpretable comparison between different machine learning models, while maintaining a lightweight architecture suitable for real-time cultural learning applications. In small-scale gesture datasets such as the one considered in this study, classical machine learning models based on geometric features often provide more stable training behavior than deep neural networks, which typically require larger datasets to avoid overfitting. For this reason, three widely adopted baseline supervised classification algorithms for structured feature-based gesture recognition were selected: Random Forest (RF), Light Gradient Boosting Machine (LightGBM), and Support Vector Machine (SVM) with radial basis function kernel. These models represent complementary learning paradigms, including bagging-based ensemble learning, gradient boosting decision trees, and margin-based classifiers, and are commonly used in gesture recognition and human pose classification tasks when datasets are relatively limited in size. The algorithms were trained using the scikit-learn library (version 1.3.2) using identical feature representations consisting of 20 geometric distance features derived from the reconstructed 3D hand landmarks and data splits to ensure full comparability across models, and were evaluated within the same experimental framework.

The hyperparameters of all classifiers were selected with the goal of balancing recognition performance, computational efficiency, and robustness, taking into account the limited size of the available dataset. Hyperparameter values were determined through preliminary empirical testing and validation experiments, evaluating different configurations and selecting those providing the most stable performance. For the RF classifier, the number of trees was fixed to 100, with a random seed fixed to 42 to ensure reproducibility. Additional tuning of parameters such as maximum tree depth, minimum number of samples per leaf, and feature selection strategy (using $max_features = \text{“}sqrt\text{”}$) did not produce measurable performance gains; therefore, default values were retained. The Gini impurity criterion

and bootstrap sampling were used for tree construction. The LightGBM classifier was implemented using 200 estimators and a fixed random seed of 42, with regularization parameters set to $\alpha = 0.5$ and $\gamma = 0.5$. The learning rate was set to the default value of 0.1. These values were selected to mitigate overfitting and avoid excessive model complexity, given the reduced dataset size. Among the evaluated classifiers, LightGBM was the only model that consistently benefited from probability calibration. Calibration was therefore applied using a sigmoid method with five-fold cross-validation, implemented through the *CalibratedClassifierCV* procedure. This configuration provided the best compromise between training time and predictive performance, leading to improved class probability estimates and more reliable decision boundaries. For the SVM classifier, the regularization parameter C and the kernel parameter γ were tuned on the validation set, exploring values $C \in [0.1, 1, 10, 50]$ and $\gamma \in [scale, 0.1, 0.01, 0.001]$. The best configuration, selected according to the macro-averaged F1-score, resulted in $C = 50$ and $\gamma = 0.01$.

To improve robustness against ambiguous or unseen gestures, an Isolation Forest (IF) was integrated into the pipeline as an outlier detection module. The IF was trained solely on the training feature vectors and used at inference time to identify samples that deviated significantly from the learned data distribution. Samples flagged as outliers were directly assigned to the *Others* class, regardless of the prediction produced by the supervised classifier. At inference time, a new RGB–D frame is first processed to extract the geometric feature vector, which is subsequently standardized using the training statistics. The standardized feature vector is then evaluated by the Isolation Forest to determine whether it lies within the known feature distribution. If classified as an inlier, the final gesture label is produced by the supervised classifier; otherwise, the sample is rejected and assigned to the *Others* class. The IF was configured by automatically selecting the contamination parameter through tuning on the validation set. Candidate contamination values were tested in the range $[0.01, 0.025, 0.05, 0.075, 0.1]$, and the optimal value was chosen based on the highest macro-averaged F1-score, ensuring an effective trade-off between rejection of outliers and preservation of valid samples.

Final models' performance was assessed through multiple quantitative metrics, comprising macro-averaged accuracy, precision, recall, and F1-score.

2.4. Virtual Reality Application

A non-immersive VR application was developed using Unity3D (Unity Technologies, San Francisco, CA, USA, version 2022.3.61f1) to implement an interactive digital totem conceived for public arrival spaces, such as airport terminals. The application simulates an airport terminal environment in which users can approach a dedicated interactive totem designed to support cultural learning and engagement. This scenario was selected to reflect realistic points of first contact for foreign visitors, where exposure to culturally specific non-verbal communication can facilitate orientation, curiosity, and intercultural understanding.

The overall system architecture consists of three main components: (i) the Unity-based front-end responsible for rendering the virtual environment and managing user interaction, (ii) an RGB–D sensing module used for hand acquisition, and (iii) a machine learning recognition module implemented in Python. These components communicate through a client–server architecture that enables real-time gesture recognition while keeping the interactive application responsive.

Figure 4 shows the workflow followed inside the Unity application.

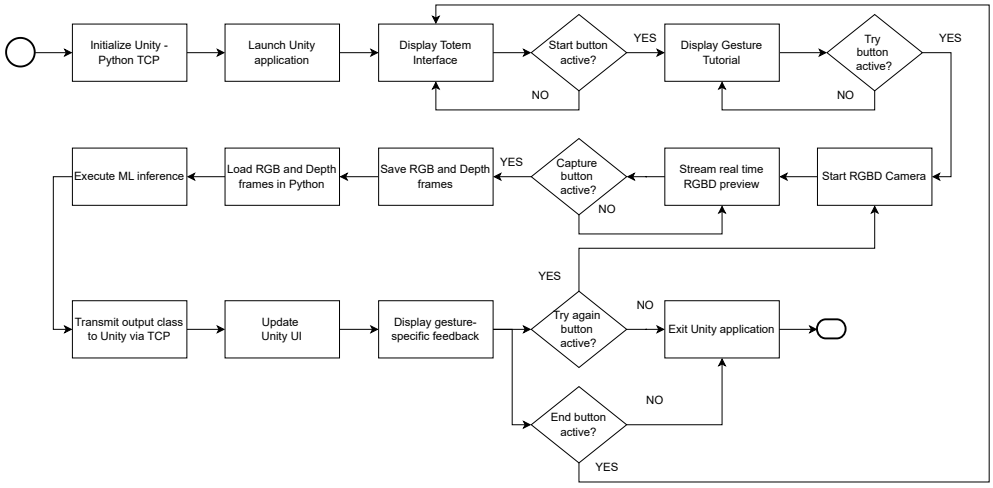


Figure 4. Unity application activity workflow.

The system is presented as a playful and interactive experience, while maintaining a clear educational purpose focused on gesture learning. When the user positions themselves in front of the Unity interactive totem, a contextual interface is activated. The interaction process is structured in three main phases: an instructional phase, a gesture reproduction phase, and a recognition feedback phase. During the instructional phase, the user can access short tutorial videos illustrating the four target gestures and explaining their cultural meaning. Once the user decides to attempt a gesture, the system activates the gesture acquisition mode, capturing frames from the RGB–D camera. Gesture recognition is implemented through a client–server architecture that enables real-time communication between the Unity application and a machine learning processing module via TCP (Transmission Control Protocol) sockets. The recognition pipeline operates as follows:

1. Unity captures a synchronized RGB–D frame from the camera and transmits the image data to the Python processing module.
2. The Python module extracts hand landmarks using MediaPipe, reconstructs their three-dimensional coordinates, computes the geometric distance features, and applies the pre-trained machine learning classifier to infer the gesture class and associated confidence score.
3. The predicted gesture label and confidence value are sent back to the Unity application through the TCP connection.
4. The recognition result is visualized directly on the totem interface within the virtual airport environment, providing immediate feedback to the user and indicating whether the gesture has been correctly recognized.

This architecture enables the seamless integration of real-time gesture recognition within the interactive learning experience, allowing users to actively reproduce gestures and immediately observe the system's interpretation of their actions. The modular separation between the Unity front-end and the machine learning back-end also facilitates future extensions of the system, allowing the recognition module to be reused in other interactive cultural installations. An example of the interactive visualization in the VR application can be seen in Figure 5.

Figure 5. Screenshots of the developed Unity-based application simulating an interactive airport information totem for learning Italian hand gestures. The figure illustrates the complete interaction workflow, from instructional content presentation to user gesture performance and real-time automatic recognition with visual feedback. (**a**) Home interface of the virtual totem. (**b**) Tutorial interface for gesture learning. (**c**) Single gesture instructional view. (**d**) Horns instructional view. (**e**) Crossed fingers instructional view. (**f**) Perfection instructional view. (**g**) User performing a gesture in front of the totem. (**h**) Real-time gesture recognition and feedback display.

3. Results

The proposed approach was evaluated at two complementary levels. Section 3.1 reports the quantitative evaluation of the gesture recognition models in terms of classification performance, while Section 3.2 presents the user-based evaluation of the developed application, focusing on usability, engagement, and learning effectiveness.

3.1. Model Performance Evaluation

The proposed gesture recognition framework was evaluated to assess both classification performance and robustness across gesture classes. Three supervised classifiers (RF, LightGBM, and SVM) were trained using the same feature representation and preprocessing pipeline, and an Isolation Forest (IF) module was additionally integrated to

investigate its effect on anomaly handling. To ensure reliable performance estimation given the limited dataset size, a 5-fold cross-validation strategy was adopted, allowing the assessment of both the average performance and the variability across folds. An ablation analysis was also conducted by comparing each classifier with and without IF. Table 3 reports the macro-averaged classification performance of all evaluated configurations in terms of mean value and standard deviation.

Table 3. Comparison of classification models using macro-averaged performance metrics under a 5-fold cross-validation protocol. Results are reported as mean ± standard deviation.

Model	Accuracy	F1 Score	Precision	Recall
RF	0.823 ± 0.019	0.819 ± 0.020	0.834 ± 0.024	0.821 ± 0.018
RF + IF	0.819 ± 0.019	0.816 ± 0.018	0.835 ± 0.019	0.818 ± 0.018
LightGBM	0.781 ± 0.026	0.779 ± 0.028	0.791 ± 0.031	0.781 ± 0.026
LightGBM + IF	0.781 ± 0.026	0.779 ± 0.028	0.791 ± 0.031	0.781 ± 0.026
SVM	0.645 ± 0.040	0.644 ± 0.042	0.661 ± 0.050	0.645 ± 0.042
SVM + IF	0.649 ± 0.033	0.647 ± 0.038	0.665 ± 0.046	0.649 ± 0.034

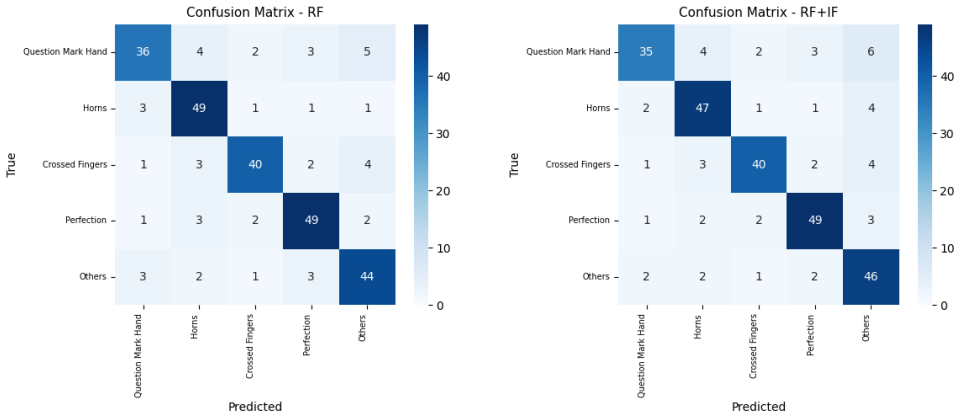
RF achieved the best performance (accuracy = 0.823 ± 0.019, F1-score = 0.819 ± 0.020), outperforming LightGBM and SVM. The low standard deviation across folds indicates stable and consistent behavior. The inclusion of IF produced only marginal changes in classification metrics for all models. However, this component is expected to enhance system robustness by enabling the detection of out-of-distribution hand configurations in real-world interaction scenarios. Based on these results, RF was selected as the reference model, and a more detailed analysis was then conducted by comparing RF with and without IF in order to assess the impact of anomaly detection on system behavior. Table 4 reports the per-class performance metrics computed by aggregating the predictions obtained on the test fold of each iteration of the cross-validation procedure. In this way, each sample is evaluated only once by a model that was not trained on it, enabling a comprehensive and unbiased assessment of the entire dataset.

Table 4. Class-wise performance comparison between Random Forest (RF) and RF + Isolation Forest (RF + IF) under 5-fold cross-validation.

Metric	Question Mark Hand	Horns	Crossed Fingers	Perfection	Others
F1-score (RF)	0.77	0.84	0.83	0.85	0.81
F1-score (RF + IF)	0.77	0.83	0.83	0.86	0.79
Precision (RF)	0.82	0.80	0.87	0.84	0.79
Precision (RF + IF)	0.85	0.81	0.87	0.86	0.73
Recall (RF)	0.72	0.89	0.80	0.86	0.83
Recall (RF + IF)	0.70	0.85	0.80	0.86	0.87

A class-wise comparison between the RF and RF + IF configurations reveals that both models exhibit consistent performance, with only marginal differences across all gesture categories. The *Perfection* gesture remains the most reliably recognized class in both configurations, achieving the highest and most balanced precision and recall values. Similarly, *Horns* and *Crossed Fingers* show stable performance, with only negligible variations between RF and RF + IF. The *Question Mark Hand* gesture represents the most challenging class, particularly in terms of recall, suggesting a higher degree of intra-class variability or geometric similarity with other gestures. This behavior is consistent across both configurations, indicating that the limitation is primarily related to feature separability rather than the presence of outliers. Regarding the *Others* class, the RF + IF configuration

shows a tendency toward higher recall but slightly lower precision compared to RF alone. This behavior suggests that the anomaly detection module increases the sensitivity to non-target or ambiguous configurations, at the cost of introducing a higher number of false positives. To provide a comprehensive evaluation of the model behavior, confusion matrices (Figure 6) were computed by aggregating the predictions obtained across all test folds of the cross-validation procedure.



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

Figure 6. Confusion matrices obtained under the cross-validation protocol for (**a**) Random Forest and (**b**) RF + Isolation Forest.

The comparison between the confusion matrices further confirms the quantitative results. While RF achieves slightly higher overall accuracy, the RF + IF configuration improves the recognition of the *Others* class, indicating enhanced sensitivity to ambiguous or out-of-distribution inputs. While this comes at the cost of a marginal decrease in overall accuracy, it represents a desirable trade-off in real-world applications, where robustness to unexpected inputs is critical.

Overall, the per-class analysis shows that the inclusion of IF does not significantly improve classification accuracy, but contributes to shaping the decision boundaries in a way that is beneficial for handling ambiguous or out-of-distribution inputs. For this reason, despite the slightly higher performance achieved by the standalone RF, the RF + IF configuration was selected for deployment, as it improves system robustness by reducing forced predictions in uncertain interaction scenarios.

To further support model interpretability, a SHAP (SHapley Additive exPlanations) analysis was conducted to quantify the contribution of each feature to the selected classifier's predictions. The SHAP analysis was performed using a sampling-based approach with the number of samples used for estimating Shapley values set to 500, and the link function was set to *identity*, ensuring that SHAP values directly reflect the contribution of each feature to the raw model output. The resulting SHAP summary plot, reported in Figure 7, provides a global view of feature importance and illustrates how variations in the selected landmark distance features influence the model output.

The SHAP summary plot provides insight into the contribution of individual geometric distance features to the chosen classifier's predictions. The plot ranks features according to their global importance, measured as the mean absolute SHAP value, and illustrates both the magnitude and direction of their influence on the model output. The most relevant features are primarily distances between metacarpophalangeal (MCP) joints of adjacent fingers, such as MIDDLE_MCP_RING_MCP, INDEX_MCP_INDEX_PIP, and INDEX_MCP_MIDDLE_MCP. This indicates that the relative spacing between fingers plays a dominant role in distinguishing between the considered gestures, which is consis-

tent with their semantic definitions, as many Italian gestures are characterized by specific finger grouping and separation patterns. Distances related to intra-finger articulation, such as PINKY_DIP_PINKY_TIP and RING_PIP_RING_DIP, also exhibit a significant impact, suggesting that the degree of finger extension or curvature contributes meaningfully to the model's decision process. In contrast, wrist-related features, including WRIST_INDEX_MCP and WRIST_THUMB_CMC, appear less influential, which is expected given the static and posture-centric nature of the selected gestures. The color distribution further reveals how feature values modulate the model output: higher distance values (red points) tend to push predictions toward specific gesture classes, while lower values (blue points) have the opposite effect, confirming the classifier reliance on anatomically meaningful spatial relationships between fingers and joints, rather than spurious correlations, and thus reinforcing the suitability of the selected geometric features for gesture recognition in educational and interactive settings.

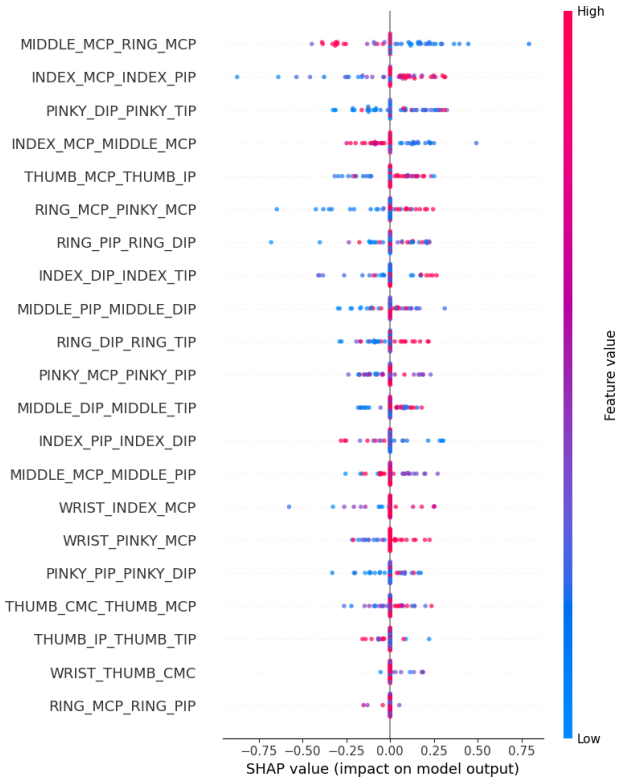


Figure 7. SHAP summary plot illustrating the contribution of the selected 3D hand landmark distance features to the selected classifier predictions. Feature values are color-coded from low to high, indicating their influence on the model output.

3.2. VR Application Evaluation

An experimental user study was conducted to evaluate the proposed Unity-based interactive totem in terms of usability, user engagement, and learning effectiveness. The study simulated a realistic scenario in which first-time users encounter culturally specific Italian gestures in a public context (e.g., an airport terminal).

A total of 25 participants (age range: 20–37) were recruited on a voluntary basis. Given the focus of the study on cultural learning, participants were selected among non-native Italian speakers and individuals with limited familiarity with Italian nonverbal communication. Each session lasted approximately 15 min and was conducted in a controlled indoor environment. The system was deployed on a standard desktop setup (Intel i7 CPU, 16 GB RAM, NVIDIA RTX 3060 GPU) with an Intel RealSense SR305 RGB–D camera. This configuration is comparable to common multimedia kiosks, supporting the feasibility

of real-world deployment. All participants provided informed consent, and the study followed standard ethical guidelines.

At the beginning of the session, each participant received a brief explanation of the study objectives and the interaction context, and was then asked to complete an initial pre-test questionnaire aimed at assessing their baseline knowledge of the semantic meaning associated with the selected Italian hand gestures. Following the pre-test, participants were allowed to freely interact with the interactive totem application. During this phase, they could explore the tutorial content and observe example videos of the gestures, and actively perform the gestures in front of the system to trigger real-time recognition and feedback. No time constraints were imposed to preserve a natural interaction. After the interaction, participants completed a post-test identical to the pre-test, enabling the evaluation of learning gains attributable to the use of the system, along with usability and engagement questionnaires.

Usability was assessed using the System Usability Scale (SUS) [36], a standardized 10-item questionnaire rated on a five-point Likert scale. Scores were computed on a 0–100 scale following Brooke [37], where values above 68 indicate above-average usability [38]. Table 5 reports the item-level results of the SUS questionnaire, summarized using the median, minimum, and maximum values across participants.

Table 5. Questions for the SUS assessment user testing. The median, minimum, and maximum values summarize the questionnaire results.

Item #	Questionnaire Item	Min	Median	Max
1	I think that I would like to use this system frequently	3	4	5
2	I found the system unnecessarily complex	1	1	3
3	I thought the system was easy to use	2	5	5
4	I think that I would need the support of a technical person to be able to use this system	1	2	5
5	I found the various functions in this system were well integrated	2	4	5
6	I thought there was too much inconsistency in this system	1	2	4
7	I would imagine that most people would learn to use this system very quickly	2	5	5
8	I found the system very cumbersome to use	1	1	4
9	I felt very confident using the system	2	5	5
10	I needed to learn a lot of things before I could get going with this system	1	1	5

The SUS results indicate high perceived usability, with a median score of 82.5, well above the standard threshold of 68. The interquartile range (Q1 = 68.75, Q3 = 93.75) suggests generally consistent evaluations, despite some variability (minimum = 40). Item-level analysis confirms strong usability perception, with positive items related to ease of use, learnability, and confidence showing high median scores. Conversely, negatively worded items (complexity, inconsistency, need for support) received low scores, indicating that the system was not perceived as difficult to use. While overall SUS scores indicate high perceived usability, certain items exhibit greater variability, particularly those related to perceived complexity, need for technical support, and system cumbersomeness. Although median values for these negatively worded items remained low, indicating general agreement on ease of use, the broader response range suggests heterogeneity in participants' feedback, reflecting individual differences in technological confidence or momentary interaction uncertainty.

User engagement was evaluated using the short form of the User Engagement Scale (UES-SF) [39], which measures engagement across four dimensions: Focused Attention (FA), Perceived Usability (PU), Aesthetic Appeal (AE), and Reward (RW). Table 6 reports the item-level results of the UES questionnaire, summarized using the median, minimum, and maximum values across participants.

Table 6. Questions for the UES assessment's user testing. The median, minimum, and maximum values are used to display the overall results of the questionnaire.

Item #	Questionnaire Item	Min	Median	Max
FA-1	I lost myself in this experience	1	3	5
FA-2	The time I spent using the Application just slipped away	1	3	5
FA-3	I was absorbed in this experience	1	4	5
PU-1	I felt frustrated while using this Application	1	1	4
PU-2	I found this Application confusing to use	1	1	4
PU-3	Using this Application was taxing	1	1	3
AE-1	This Application was attractive	2	4	5
AE-2	This Application was aesthetically appealing	2	4	5
AE-3	This Application appealed to my senses	3	4	5
RW-1	Using Application was worthwhile	1	4	5
RW-2	My experience was rewarding	1	5	5
RW-3	I felt interested in this experience	2	5	5

The UES revealed a high overall engagement level, obtaining a median value of 4.02, an IQR of 0.75, a minimum value of 2.83, and a maximum value equal to 5.00. Among the subscales, RW achieved the highest median score (4.67 with IQR equal to 1.00). RW-related items achieved the highest central tendency overall, particularly "I felt interested in this experience", which approached ceiling values, underscoring the motivational and cultural relevance of the application. PU also showed high central tendency (median = 4.33, IQR = 1.00), consistent with the SUS findings. AE obtained a median of 4.13 with an IQR of 1.67, reflecting positive but somewhat heterogeneous aesthetic perception. FA presented lower central tendency (median = 3.00, IQR = 1.17), with greater dispersion, suggesting variability in the degree of immersion experienced by participants. Specifically, the item "I lost myself in this experience" displayed the widest response range (1–5). This variability may reflect differences in the interpretation of immersion-related terminology rather than reduced engagement. In contrast, the item "I was absorbed in this experience" showed higher central tendency and reduced dispersion, indicating sustained attention when described in more direct terms. Overall, the engagement profile indicates that the application is perceived as usable, visually appealing, and rewarding, while eliciting moderate but meaningful levels of attentional absorption.

To evaluate the effectiveness of the application as a learning tool, a dedicated short-term learning effect assessment was conducted using a pre–post open-answer questionnaire design. Participants were asked to complete a short questionnaire before and after interacting with the system, aimed at assessing their understanding of the semantic meaning associated with the selected Italian hand gestures, requiring participants to identify or describe the meaning conveyed by each gesture. Pre-interaction responses provided a baseline measure of prior knowledge, while post-interaction responses captured knowledge acquired through the use of the interactive experience. Differences between pre- and post-test responses were used as an indicator of learning gain attributable to the application.

To assess the learning effect of the system, the responses collected through the pre-post questionnaire were statistically analyzed. Each answer was manually coded as correct or incorrect based on the expected meaning of the gesture. A total recognition score

ranging from 0 to 4 was computed for each participant by summing the number of correctly identified gestures.

Before the experience, correct identification rates were generally low across all gestures. A substantial proportion of participants explicitly declared that they did not know the meaning of the gestures (12 out of 25 for *Question Mark Hand*, 13 out of 25 for *Horns*, 8 out of 25 for *Crossed Fingers*, and 17 out of 25 for *Perfection*). For the *Horns* gesture, 7 out of 25 participants associated it with "rock and roll", reflecting a widespread Anglo-American interpretation rather than its culturally specific Italian meaning. This misinterpretation was among the most recurrent patterns in the pre-test responses. For the *Question Mark Hand*, 3 out of 25 participants explicitly associated the gesture with "mamma mia", a common stereotype linked to Italian expressiveness rather than to its pragmatic communicative function. This suggests that participants relied on media-driven cultural imagery rather than semantic understanding. Similarly, the *Perfection* gesture was frequently confused with the internationally recognized "OK" sign, and only a very limited number of participants correctly identified its culturally specific meaning related to precision or careful execution. Overall, the pre-test results indicate that although some gestures were visually familiar, their communicative meanings were largely unknown or inaccurately interpreted.

After interacting with the system, correct response rates increased substantially across all gestures, with only a few subjects declaring that they did not recall the meaning of the specific gesture (none for the *Question Mark Hand*, 3 out of 25 for *Horns*, 2 out of 25 for *Crossed Fingers*, 3 out of 25 for *Perfection*). Nearly all participants correctly identified the meaning of the *Question Mark Hand* as expressing questioning, disbelief, or "What do you want?", "What do you mean?", "What are you saying?". However, minor residual misconceptions remained: two participants interpreted it as "perfect", and one associated it again with "mamma mia", indicating that stereotypical interpretations were not entirely eliminated. For the *Horns* gesture, the association with "rock and roll" completely disappeared, and the majority of responses correctly referred to protection against bad luck or warding off negative energy. The *Crossed Fingers* gesture showed a marked consolidation of meaning, with most participants correctly identifying it as a sign of good luck or hope. The *Perfection* gesture, which had been one of the least understood in the pre-test phase, was predominantly interpreted after the experience as meaning "be careful", "do it properly", or "in the right way", demonstrating acquisition of its culturally specific nuance.

Differences between pre-test and post-test scores were analyzed using the Wilcoxon signed-rank test, which is appropriate for paired non-parametric data. Statistical significance was evaluated at a significance level of $p < 0.05$. The results indicate an improvement in participants' ability to correctly identify the gestures after interacting with the system. The median recognition score increased from the pre-test to the post-test. The analysis using the Wilcoxon signed rank test revealed a statistically significant performance improvement ($W = 10$, $Z = -3.80$, $p < 0.001$), indicating a measurable short-term learning effect. The effect size was large (r = 0.81), suggesting a substantial improvement in participants' ability to correctly identify the gestures after the experience. In addition to the significant increase in the total recognition score, McNemar's exact tests were conducted for each individual gesture to assess changes in the proportion of correct responses between pre-test and post-test conditions. The results, displayed in Table 7, showed statistically significant improvements for all gestures: Question Mark Hand ($p = 0.0039$), Horns ($p = 0.0000153$), Crossed Fingers ($p = 0.0391$), and Perfection ($p = 0.0005$).

Table 7. Pre–post comparison of gesture meaning recognition. Percentages refer to the proportion of correct responses among the 25 participants. Statistical significance was assessed using McNemar's exact test.

Gesture	Correct Pre	Correct Post	Improvement	*p*-Value
Question Mark Hand	13 (52%)	22 (88%)	+36%	0.0039
Horns	5 (20%)	22 (88%)	+68%	0.0000153
Crossed Fingers	16 (64%)	23 (92%)	+28%	0.0391
Perfection	8 (32%)	22 (88%)	+56%	0.0005

Overall, the results indicate a clear short-term learning effect. The system not only improved recognition accuracy but also reduced culturally inaccurate interpretations, supporting the acquisition of gesture meanings through interactive and embodied learning.

4. Discussion

This work introduced a machine learning–based framework for the recognition of culturally specific Italian hand gestures, conceived as a tool to support the learning and dissemination of intangible cultural heritage through interactive technologies. By combining RGB–D sensing, geometric feature extraction, interpretable classification models, and integration within a virtual environment, the approach aims to bridge automatic gesture recognition with cultural education contexts. The framework contributes to the literature by demonstrating that lightweight, interpretable machine learning approaches can effectively support culturally grounded gesture recognition without requiring large-scale datasets or complex deep learning architectures. From a cultural heritage perspective, the system highlights the potential of interactive technologies to support the transmission of intangible cultural elements, such as gestures, which are often difficult to formalize and preserve. By enabling real-time recognition and feedback within an immersive environment, it opens new opportunities for engaging and accessible cultural learning experiences.

The comparative evaluation of three supervised classifiers showed that the Random Forest model achieved the best overall performance (macro-averaged accuracy and F1-score of 0.82), surpassing LightGBM and SVM. The additional analysis, including the IF module, highlighted that anomaly detection does not significantly improve classification accuracy, but contributes to shaping the decision boundaries in a way that is beneficial for handling ambiguous or out-of-distribution inputs. For this reason, RF integration with IF was considered particularly relevant for deployment in real-world interactive scenarios, where user behavior may be highly variable. Class-wise evaluation highlights both strengths and limitations of the proposed feature representation. The high recall and F1-score obtained for the *Perfection* class indicate that its spatial configuration generates a distinctive signature within the selected feature space, enabling reliable discrimination. Likewise, the *Horns* class exhibits high precision, indicating that false positive predictions for this class are rare. In contrast, the reduced recall observed for the *Question Mark Hand* class suggests residual ambiguity, likely arising from partial overlap in finger grouping patterns with other gestures, as also reflected in the confusion matrix, where misclassifications predominantly occur among gestures with similar relative finger spacing. These findings point to an inherent limitation of static geometric descriptors: while effective at capturing global hand posture, they are less sensitive to subtle articulation details such as finger curvature or dynamic motion.

When contextualized within the broader gesture recognition literature, the obtained performance is consistent with results reported in studies employing geometric descriptors and classical machine learning models on limited datasets. The achieved macro-averaged accuracy compares favorably with studies targeting culture-specific gestures, which typ-

ically operate under constrained vocabularies and modest training data. For instance, gesture recognition applied to classical dance mudras using MediaPipe-based pipelines reported similar feasibility outcomes, confirming that culturally grounded gesture modeling remains viable even without large-scale training data [34]. Comparable trends have also been observed in the recognition of intangible cultural heritage dance movements using hybrid feature representations, where performance gains were primarily driven by feature fusion rather than model complexity [33]. These observations suggest that, for domain-specific cultural gesture recognition under limited-data conditions, carefully designed geometric descriptors combined with ensemble learning can provide effective performance while maintaining computational efficiency. Nevertheless, future work with larger datasets could explore the integration of deep learning architectures to further investigate potential performance improvements. Moreover, the interpretability offered by SHAP analysis represents an additional contribution rarely emphasized in cultural heritage gesture recognition studies, strengthening transparency and facilitating integration within educational contexts.

An additional aspect of the proposed pipeline is the integration of depth information during feature extraction. By replacing the estimated landmark depth provided by MediaPipe with the depth values measured by the RGB–D sensor, the system reconstructs true three-dimensional landmark coordinates and computes physically consistent inter-joint distances. This RGB–D integration provides a more stable geometric representation of hand posture across variations in viewpoint, hand orientation, and lighting conditions. Although a direct RGB-only comparison was not performed in the present exploratory study, the use of depth sensing contributes to a more reliable description of hand configuration and supports the robustness of the proposed feature representation.

Beyond classification performance, the proposed framework was integrated into an interactive virtual environment designed to support experiential learning of culturally specific gestures, demonstrating the feasibility of embedding lightweight recognition pipelines within real-time applications. The results of the user-based evaluation indicate that participants perceived the system as usable and engaging, as reflected by favorable SUS and UES scores. In particular, the high ratings associated with PU and RW dimensions suggest that the interaction paradigm was accessible and motivating, supporting sustained participation in the learning activity. The FA dimension showed comparatively lower median values and greater dispersion. However, item-level inspection indicates that this variability was largely driven by the statement “I lost myself in this experience” which may be interpreted ambiguously. When immersion was framed in more direct terms (“I was absorbed in this experience”), ratings were notably higher. These findings suggest that the application achieves meaningful attentional engagement appropriate to its desktop, non-immersive configuration.

The statistical analysis on the data obtained from the pre–post assessment of gesture comprehension supports the presence of a short-term learning effect. The significant increase in the overall recognition score, confirmed by the Wilcoxon signed-rank test, suggests that the interactive system can effectively support users in understanding. The statistical analysis further confirms that the observed learning effect was not limited to the overall recognition score, but was also evident at the level of individual gestures. The pre-interaction responses reveal that, although the gestures were visually familiar to many participants, their meanings were largely unknown, stereotyped, or misinterpreted. In particular, the frequent association of the *Horns* gesture with “rock and roll” reflects the dominance of globally disseminated interpretations over culturally specific Italian meanings. Similarly, the repeated association of the *Question Mark Hand* with “mamma mia” indicates reliance on media-driven stereotypes rather than pragmatic understanding.

These findings suggest that gestures can be widely recognized as markers of Italian-ness at a symbolic level, while their pragmatic and communicative meanings remain poorly understood, which is an issue that is central to the transmission of intangible cultural heritage. Following interaction with the system, correct identification rates increased substantially across all gestures, and stereotypical or globally dominant misinterpretations were largely eliminated, with the improvement not limited to generic recognition but extended to contextually appropriate meanings. Although minor residual misconceptions persisted, such as two post-test responses interpreting the *Question Mark Hand* as "perfect" and one reiterating "mamma mia", the overall shift from uncertainty and stereotype to semantic clarity indicates a measurable short-term learning effect.

From a broader perspective, the integration of gesture recognition within an interactive VR environment aligns with ongoing research exploring digital technologies as tools for safeguarding and disseminating intangible cultural heritage [40,41]. These works typically emphasize immersion and narrative engagement but often rely on pre-recorded content or scripted interaction rather than real-time behavioral interpretation [10,42]. The learning outcomes findings complement the engagement findings, as the high RW scores observed in the UES indicate that participants also perceived the experience as worthwhile and motivating, an aspect that is particularly relevant for cultural heritage dissemination, where sustained interest is essential for meaningful transmission. Moreover, the system's lightweight architecture and deployment on consumer hardware make it suitable for installation in public spaces such as airports, museums, or cultural events, where attentional engagement, combined with high usability and motivational appeal, may be more desirable than fully immersive experiences requiring specialized equipment.

Despite encouraging performance, some limitations of the work should be acknowledged. First, the dataset size remains relatively modest, which restricts the diversity of hand shapes, execution styles, and environmental variability captured during training. While this dataset was sufficient to conduct the exploratory evaluation presented in this work, the limited number of participants restricts the variability captured during training. In particular, the dataset does not fully represent the diversity of hand morphology, skin tones, gesture execution styles, and environmental conditions such as lighting and viewpoint variations. Although cross-validation was adopted to mitigate this issue, future work will focus on expanding the dataset to include a larger and more diverse group of participants and acquisition conditions in order to improve the robustness and external validity of the gesture recognition system. This limited data availability also influenced the methodological choices adopted in this study. In particular, deep learning architectures, which typically require large-scale datasets to achieve stable training and avoid overfitting, were not evaluated in the present work. Instead, the study focused on lightweight machine learning models operating on structured geometric descriptors, which are more suitable for small-data scenarios and enable interpretable and computationally efficient recognition pipelines. Future work will therefore focus on expanding the dataset in order to investigate the potential benefits of deep learning approaches for gesture recognition in this cultural heritage context. Second, the reliance on static geometric descriptors inherently limits sensitivity to subtle articulation nuances and dynamic motion cues, which may contribute to the observed confusion between visually similar gestures. Incorporating temporal information or joint-angle representations may improve separability in future iterations. Third, the learning assessment was limited to immediate post-interaction responses. Although the pre–post comparison clearly demonstrates short-term knowledge acquisition, the study does not evaluate long-term retention. Future investigations should include delayed post-tests to determine whether gesture meanings are consolidated over time. Finally, the current implementation focuses on four gestures representative of Italian

cultural expression. Although this targeted approach ensures depth and clarity, it does not address the broader variability of Italian gestural communication across regions and contexts. Extending the framework to a larger and more diverse gesture set would strengthen its contribution to intangible cultural heritage preservation.

5. Conclusions

This paper presented a machine learning–based framework for the recognition and learning of iconic Italian hand gestures, conceived as a technological tool for the dissemination of intangible cultural heritage and intercultural communication. By combining RGB–D sensing, geometric feature extraction, and an interpretable classification pipeline, the proposed system enables reliable recognition of culturally grounded Italian gestures while maintaining low computational complexity and transparency. Experimental results obtained through a 5-fold cross-validation procedure demonstrated that the proposed approach achieves solid and consistent recognition performance, with a macro-averaged accuracy and F1-score of 0.82, despite the limited dataset size and the subtle similarities between gesture classes. The integration of an outlier detection mechanism further improved robustness, allowing the system to effectively reject non-target gestures and supporting real-world deployment scenarios. The framework was embedded into an interactive virtual environment designed to support playful and accessible learning experiences. The user study confirmed high levels of perceived usability and engagement, and a measurable learning effect, suggesting that automatic gesture recognition can play a meaningful role in facilitating the understanding of culturally specific nonverbal communication.

Overall, this work demonstrates how lightweight and interpretable machine learning methods, when combined with interactive environments, can contribute to the preservation and transmission of intangible cultural practices. Future work will focus on expanding the gesture vocabulary, increasing dataset diversity across participants and environmental conditions, integrating temporal and multimodal features to improve recognition robustness, and evaluating long-term learning retention as well as real-world deployment scenarios.

Author Contributions: Conceptualization, G.D.P., I.L., A.M. and M.V.; methodology, C.I., G.D.P., I.L., A.M., M.V., S.M. and G.M.; software, G.D.P., I.L., A.M. and M.V.; validation, C.I., G.M., S.M., E.V. and L.U.; formal analysis, C.I., G.M., S.M., E.V. and L.U.; investigation, C.I., G.D.P., I.L., A.M., M.V. and S.M.; resources, C.I., G.M., E.V. and L.U.; data curation, C.I., G.D.P., I.L., A.M., M.V. and S.M.; writing—original draft preparation, C.I., G.D.P., I.L., A.M., M.V. and S.M.; writing—review and editing, G.M., E.V. and L.U.; visualization, G.D.P., I.L., A.M. and M.V.; supervision, G.M., E.V. and L.U.; project administration, E.V. and L.U.; funding acquisition, E.V. and L.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical committee approval was not required, as the study involved non-invasive, observational procedures only, no medical or psychological interventions, no vulnerable populations, and no collection of sensitive personal data. Although facial images may be present in the raw recordings, facial data were not used for identification or biometric purposes, and all data were anonymized prior to analysis. The recordings were used exclusively for hand landmark and depth feature extraction.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Dataset available on request from the authors.

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

References

1. Peltier, I.; Mccafferty, S. Gesture and Identity in the Teaching and Learning of Italian. *Mind Cult. Act.* **2010**, *17*, 331–349. [CrossRef]
2. Graham, J.A.; Argyle, M. A cross-cultural study of the communication of extra-verbal meaning by gestures. *Int. J. Psychol.* **1975**, *10*, 57–67. [CrossRef]
3. Poggi, I. Symbolic Gestures: The Case of the Italian Gestionary. *Gesture* **2002**, *2*, 71–98. [CrossRef]
4. Telmon, T. La gestualità in Italia. In *La Cultura Italiana*; Cavalli Sforza, L.L., Ed.; Lingue e linguaggi; UTET: Torino, Italy, 2009; Volume II, pp. 589–648.
5. Kendon, A. Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *J. Pragmat.* **1995**, *23*, 247–279. [CrossRef]
6. Graziano, M.; Gullberg, M. Providing evidence for a well-worn stereotype: Italians and Swedes do gesture differently. *Front. Commun.* **2024**, *9*, 1314120. [CrossRef]
7. Diadori, P. Nonverbal Communication in Classroom Interaction and Its Role in Italian Foreign Language Teaching and Learning. *Languages* **2024**, *9*, 164. [CrossRef]
8. Broccolini, A. *Italian Intangible Communities: Heritage, Participation, and Identity in Contemporary Italy*; Edizioni Ca' Foscari: Venezia, Italy, 2017.
9. Innocente, C.; Ulrich, L.; Moos, S.; Vezzetti, E. A framework study on the use of immersive XR technologies in the cultural heritage domain. *J. Cult. Herit.* **2023**, *62*, 268–283. [CrossRef]
10. Innocente, C.; Nonis, F.; Lo Faro, A.; Ruggieri, R.; Ulrich, L.; Vezzetti, E. A Metaverse Platform for Preserving and Promoting Intangible Cultural Heritage. *Appl. Sci.* **2024**, *14*, 3426. [CrossRef]
11. Zhang, J.; Wan Yahaya, W.A.J.; Sanmugam, M. The Impact of Immersive Technologies on Cultural Heritage: A Bibliometric Study of VR, AR, and MR Applications. *Sustainability* **2024**, *16*, 6446. [CrossRef]
12. Ariya, P.; Wongwan, N.; Worragin, P.; Intawong, K.; Puritat, K. Immersive realities in museums: Evaluating the impact of VR, VR360, and MR on visitor presence, engagement and motivation. *Virtual Real.* **2025**, *29*, 117. [CrossRef]
13. Fiorucci, M.; Khoroshiltseva, M.; Pontil, M.; Traviglia, A.; Del Bue, A.; James, S. Machine Learning for Cultural Heritage: A Survey. *Pattern Recognit. Lett.* **2020**, *133*, 102–108. [CrossRef]
14. Pavlovic, V.; Sharma, R.; Huang, T. Visual interpretation of hand gestures for human–computer interaction: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 677–695. [CrossRef]
15. Starner, T.; Weaver, J.; Pentland, A. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1371–1375. [CrossRef]
16. Wilson, A.; Bobick, A. Parametric hidden Markov models for gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 884–900. [CrossRef]
17. Gavrila, D. The Visual Analysis of Human Movement: A Survey. *Comput. Vis. Image Underst.* **1999**, *73*, 82–98. [CrossRef]
18. Mitra, S.; Acharya, T. Gesture Recognition: A Survey. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2007**, *37*, 311–324. [CrossRef]
19. Dipietro, L.; Sabatini, A.M.; Dario, P. A Survey of Glove-Based Systems and Their Applications. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2008**, *38*, 461–482. [CrossRef]
20. Bargellesi, N.; Carletti, M.; Cenedese, A.; Susto, G.A.; Terzi, M. A Random Forest-based Approach for Hand Gesture Recognition with Wireless Wearable Motion Capture Sensors. *IFAC-PapersOnLine* **2019**, *52*, 128–133. [CrossRef]
21. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*; IEEE: Piscataway, NJ, USA, 2011; pp. 1297–1304. [CrossRef]
22. Suarez, J.; Murphy, R.R. Hand gesture recognition with depth images: A review. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*; IEEE: Piscataway, NJ, USA, 2012; pp. 411–417. [CrossRef]
23. Oikonomidis, I.; Kyriazis, N.; Argyros, A. Efficient model-based 3D tracking of hand articulations using Kinect. In *Proceedings of the British Machine Vision Conference*; BMVA Press: Durham, UK, 2011; Volume 1. [CrossRef]
24. Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.L.; Grundmann, M. MediaPipe Hands: On-device Real-time Hand Tracking. *arXiv* **2020**, arXiv:2006.10214.
25. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 1302–1310. [CrossRef]
26. Meng, Y.; Jiang, H.; Duan, N.; Wen, H. Real-Time Hand Gesture Monitoring Model Based on MediaPipe's Registerable System. *Sensors* **2024**, *24*, 6262. [CrossRef]
27. Abdallah, M.S.; Samaan, G.H.; Wadie, A.R.; Makhmudov, F.; Cho, Y.I. Light-Weight Deep Learning Techniques with Advanced Processing for Real-Time Hand Gesture Recognition. *Sensors* **2023**, *23*, 2. [CrossRef]
28. Ulrich, L.; De Luca, A.; Miraglia, R.; Mulassano, E.; Quattrocchio, S.; Marullo, G.; Innocente, C.; Salerno, F.; Vezzetti, E. A 3D Camera-Based Approach for Real-Time Hand Configuration Recognition in Italian Sign Language. *Sensors* **2026**, *26*, 1059. [CrossRef]

29. Marullo, G.; Amati, D.; Barbarino, S.; D'Onofrio, M.; Sechi, T.; Innocente, C.; Vezzetti, E.; Ulrich, L. Three-Dimensional Vision-Based Recognition of Guitar Chords. *Comput. Music. J.* **2026**, 1–42. [CrossRef]
30. Al-Barham, M.; Sa'Aleek, A.A.; Al-Odat, M.; Hamad, G.; Al-Yaman, M.; Elnagar, A. Arabic Sign Language Recognition Using Deep Learning Models. In *2022 13th International Conference on Information and Communication Systems (ICICS)*; IEEE: Piscataway, NJ, USA, 2022; pp. 226–231. [CrossRef]
31. Sadhana, P.; Ravishankar, N.; Palaniswamy, S. Bharatanatyam Mudra Recognition Using Deep Learning and Meta-Learning Techniques. In *2024 1st International Conference on Communications and Computer Science (InCCCS)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [CrossRef]
32. Ulrich, L.; Carmassi, G.; Garelli, P.; Lo Presti, G.; Ramondetti, G.; Marullo, G.; Innocente, C.; Vezzetti, E. SIGNIFY: Leveraging Machine Learning and Gesture Recognition for Sign Language Teaching Through a Serious Game. *Future Internet* **2024**, *16*, 447. [CrossRef]
33. Yang, J.; Li, X.; Liu, W.; Shao, H.; Li, G. Research on Recognition Method of Non-Legacy Dance Action Based on Multi-Feature Fusion. *Int. J. Intell. Inf. Technol.* **2025**, *21*, 1–16. [CrossRef]
34. Kishore Kumar, A.V.; Emmadisetty, R.; Chandralekha, M.; Saleem, K. Mediapipe-Powered SVM Model for Real-Time Kuchipudi Mudras Recognition. In *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 583–588. [CrossRef]
35. Lin, Y.; Jiao, X.; Zhao, L. Detection of 3D Human Posture Based on Improved Mediapipe. *J. Comput. Commun.* **2023**, *11*, 102–121. [CrossRef]
36. Lewis, J.R. The System Usability Scale: Past, Present, and Future. *Int. J. Hum.-Comput. Interact.* **2018**, *34*, 577–590. [CrossRef]
37. Brooke, J. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation in Industry*; CRC Press: Boca Raton, FL, USA, 1996; p. 6.
38. Sauro, J.; Lewis, J.R. *Quantifying the User Experience: Practical Statistics for User Research*, 1st ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2012.
39. O'Brien, H.L.; Cairns, P.; Hall, M. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *Int. J. Hum.-Comput. Stud.* **2018**, *112*, 28–39. [CrossRef]
40. Wang, Y.; Ametefe, D.S. Immersive Technologies in Digital Museums: A Systematic Review of Virtual and Augmented Reality for Heritage Reconstruction and Visitor Experience. *PRESENCE Virtual Augment. Real.* **2026**, *35*, 71–93. [CrossRef]
41. Wang, H.; Du, J.; Li, Y.; Zhang, L.; Li, X. Grand Challenges in Immersive Technologies for Cultural Heritage. *Int. J. Hum.-Comput. Interact.* **2025**, *41*, 13682–13703. [CrossRef]
42. Lin, C.; Xia, G.; Nickpour, F.; Chen, Y. A review of emotional design in extended reality for the preservation of culture heritage. *npj Herit. Sci.* **2025**, *13*, 86. [CrossRef]