applied sciences

MDPI

Article

Enhanced Cleft Lip and Palate Classification Using SigLIP 2: A Comparative Study with Vision Transformers and Siamese Networks

Oraphan Nantha, Benjaporn Sathanarugsawait and Prasong Praneetpolgrang *

School of Information Technology, Sripatum University, Bangkok 10900, Thailand; oraphan.nan@spumail.net (O.N.); benjaporn.sa@spu.ac.th (B.S.)
* Correspondence: prasong.pr@spu.ac.th

Abstract: This paper extends our previous work on cleft lip and/or palate (CL/P) classification, which employed vision transformers (ViTs) and Siamese neural networks. We now integrate SigLIP 2, a state-of-the-art multilingual vision–language model, for feature extraction, replacing the previously utilized BiomedCLIP. SigLIP 2 offers enhanced semantic understanding, improved localization capabilities, and multilingual support, potentially leading to more robust feature representations for CL/P classification. We hypothesize that SigLIP 2's superior feature extraction will improve the classification accuracy of CL/P types (bilateral, unilateral, and palate-only) from the UltraSuite CLEFT dataset, a collection of ultrasound video sequences capturing tongue movements during speech with synchronized audio recordings. A comparative analysis is conducted, evaluating the performance of our original ViT-Siamese network model (using BiomedCLIP) against a new model leveraging SigLIP 2 for feature extraction. Performance is assessed using accuracy, precision, recall, and F1 score, demonstrating the impact of SigLIP 2 on CL/P classification. The new model achieves statistically significant improvements in overall accuracy (86.6% vs. 82.76%) and F1 scores for all cleft types. We discuss the computational efficiency and practical implications of employing SigLIP 2 in a clinical setting, highlighting its potential for earlier and more accurate diagnosis, personalized treatment planning, and broader applicability across diverse populations. The results demonstrate the significant potential of advanced vision–language models, such as SigLIP 2, to enhance AI-powered medical diagnostics.

Keywords: cleft lip and palate; vision–language models; few-shot learning; medical image analysis; AI in healthcare

Academic Editors: Gang Wei and Pedro Couto

Received: 24 February 2025
Revised: 29 March 2025
Accepted: 19 April 2025
Published: 25 April 2025

Citation: Nantha, O.; Sathanarugsawait, B.; Praneetpolgrang, P. Enhanced Cleft Lip and Palate Classification Using SigLIP 2: A Comparative Study with Vision Transformers and Siamese Networks. *Appl. Sci.* **2025**, *15*, 4766. https://doi.org/10.3390/app15094766

1. Introduction

Cleft lip and/or palate (CL/P) is a common congenital anomaly, affecting approximately one in every 700 to 1000 births worldwide [1]. This condition results from incomplete fusion of facial structures during fetal development, leading to a split or opening in the upper lip and/or palate. CL/P severity varies widely, from a minor lip notch to a complete bilateral separation of both lip and palate [1]. Beyond its impact on facial appearance, CL/P significantly affects essential functions like feeding, speech, hearing, and dental development [1]. The interplay of genetic and environmental factors contributes to CL/P's etiology, making early and accurate diagnosis crucial for effective treatment and improved long-term outcomes [1]. Timely intervention, including surgery and rehabilitative therapies, is vital to address both cosmetic and functional impairments, ultimately improving the quality of life for those affected.

Building upon our prior research [2], this study extends our investigation into CL/P classification by introducing a novel feature extraction approach. Our previous work [2] combined vision transformers (ViTs) and Siamese neural networks to analyze multimodal data from the UltraSuite CLEFT dataset [3], which includes ultrasound video sequences of tongue movements and synchronized audio recordings. In that study, ViTs captured long-range dependencies and global context within the ultrasound images and spectrograms, while Siamese networks facilitated effective few-shot learning, a critical capability given the limited labeled data in medical imaging [2,4,5]. That approach demonstrated promising results, achieving an overall classification accuracy of 82.76% across the three CL/P types: BCLP, CP, and UCLP [2]. However, a key limitation of that prior work was its reliance on BiomedCLIP [6] for feature extraction. BiomedCLIP, while effective, is primarily trained on English biomedical text, potentially limiting its ability to capture the full range of nuances in multilingual speech data [7–9] or the subtle visual details crucial for distinguishing CL/P variations [10–16].

This paper addresses that limitation by incorporating SigLIP 2 [17], a state-of-the-art multilingual vision–language encoder. SigLIP 2, building upon SigLIP [18] and models like CLIP [19], offers significant advantages. It demonstrates improved semantic understanding, capturing more nuanced relationships between visual and textual information [17,20,21]. Its enhanced localization capabilities allow for more precise identification of relevant image features, crucial for analyzing subtle anatomical variations in CL/P ultrasound images [17]. Furthermore, SigLIP 2's inherent multilingual support makes it better suited for analyzing diverse speech data, a common scenario in CL/P research [17]. SigLIP 2's architecture, with its improved training and larger model sizes, contributes to superior performance in various vision–language tasks [17,20].

We utilize the UltraSuite CLEFT dataset to evaluate our approach. This dataset, designed for CL/P research, provides multimodal data for analyzing speech production in children with cleft conditions. It includes synchronized ultrasound videos of tongue movements and audio recordings. The ultrasound videos provide visual information on tongue articulation, affected by CL/P, while the audio captures acoustic characteristics reflecting potential speech impairments. These complementary modalities, along with textual prompts, enable a comprehensive analysis of speech and articulatory movements relevant to CL/P classification.

This study addresses the following research question: Does incorporating SigLIP 2 for feature extraction improve the accuracy and efficiency of CL/P classification compared to the previous ViT-Siamese network model that utilized BiomedCLIP? We hypothesize that SigLIP 2's enhanced feature representations, stemming from its improved semantic understanding, localization capabilities, and multilingual support, will lead to a statistically significant improvement in CL/P classification performance (accuracy, precision, recall, and F1 score) compared to our previous model. This improvement is expected because SigLIP 2 can capture more nuanced and relevant information from both ultrasound images and speech spectrograms, leading to a more discriminative feature space for CL/P classification.

2. Related Works

Our previous work [2] established a foundation for cleft lip and/or palate (CL/P) classification using artificial intelligence, specifically employing vision transformers (ViTs) and Siamese neural networks. This approach was informed by several key studies. Wang et al. [10] developed a deep learning model combining LSTM and DRNN for hypernasality detection in Mandarin-speaking patients with CL/P, achieving high accuracy, albeit focusing solely on speech audio data. Zhu et al. [11] utilized a CNN framework (U-net and Dense U-net) for automatic tongue contour tracking in ultrasound images, demon-

strating the potential of deep learning for anatomical analysis in CL/P. Csapó et al. [12] explored articulatory-to-acoustic mapping using ultrasound images and residual networks, highlighting the feasibility of processing different ultrasound image representations. Al-Hammuri et al. [13] compared various segmentation techniques for tongue edge detection in ultrasound images, finding CNNs and U-nets superior to traditional methods. These studies, along with others focusing on speech assessment [14] and the psychological aspects of CL/P [15,16], underscored the need for a multimodal approach integrating both anatomical and functional information while also addressing the challenge of limited data availability in medical imaging [22]. Our previous work addressed these needs by combining ViTs and Siamese networks, achieving competitive results with few-shot learning on multimodal data [2,4,5].

Since the publication of our previous work [2], the field of vision–language models has advanced significantly. The development of SigLIP [18] and subsequently SigLIP 2 [17] represents a major step forward. SigLIP, introduced by Zhai et al. [18], proposed a novel sigmoid loss function for language–image pre-training, improving upon the contrastive loss used in models like CLIP [19]. This resulted in stronger performance on various downstream tasks. SigLIP 2 [17,20,21] further enhanced this approach, with improved training strategies, provided larger model sizes, and, crucially, provided multilingual support. This multilingual capability is particularly relevant to CL/P research, as it allows for the analysis of speech data from diverse linguistic backgrounds, broadening the applicability of AI-powered diagnostic tools. While BiomedCLIP [6] demonstrated the effectiveness of adapting vision–language models to the biomedical domain, its focus on English-language text limits its utility in multilingual contexts [7–9]. SigLIP 2's architecture and training methodology enable it to capture more nuanced semantic relationships and finer-grained visual details, making it a promising alternative for medical image analysis.

The application of vision–language models in medical imaging is a rapidly growing area of research. While direct applications of SigLIP/SigLIP 2 to CL/P are still emerging, related work demonstrates the potential of these models in other medical domains. For example, studies have explored the use of vision–language models for tasks such as medical report generation [7], disease classification from medical images [8], and visual question answering in radiology [9]. These studies highlight the ability of vision–language models to leverage both visual and textual information for improved understanding and analysis of medical data.

Few-shot learning remains a critical area of research in medical imaging, given the inherent challenges in obtaining large, labeled datasets. Recent work has explored various techniques for improving few-shot learning performance, including meta-learning approaches [4], data augmentation strategies specifically designed for medical images [5], and the use of self-supervised learning to pre-train models on unlabeled data [22]. These advancements are relevant to CL/P classification, as they offer potential avenues for further enhancing the performance of models like ours, which rely on Siamese networks for few-shot learning. The combination of advanced vision–language models like SigLIP 2 with these novel few-shot learning techniques holds significant promise for improving the accuracy and efficiency of medical image analysis, particularly in scenarios with limited labeled data.

3. Materials

This section details the materials used in this study, encompassing the dataset of ultrasound and audio recordings from children with cleft lip and/or palate (CL/P) and the SigLIP 2 model employed for feature extraction.

3.1. Dataset

This study utilizes the same UltraSuite CLEFT dataset as employed in our previous work [2]. As detailed by Eshky et al. [3], this publicly available repository contains synchronized ultrasound video sequences capturing tongue movements during speech and corresponding audio recordings. The data originate from children with various types of CL/P and typically developing children. For this classification task, we focus on the data from 29 children diagnosed with one of three CL/P types: bilateral cleft lip and palate (BCLP), cleft palate only (CP), and unilateral cleft lip and palate (UCLP). The dataset also includes specific textual prompts used during recording to elicit relevant articulatory movements. For comprehensive details regarding participant demographics, the specific recording setup and procedures, and the exact textual prompts used, readers are directed to the original dataset publication [3] and our prior work [2], which provides context specific to our classification approach.

3.2. SigLIP 2

SigLIP 2 (Sigmoid Loss for Language Image Pre-training 2) [17] is a state-of-the-art vision–language model that builds upon the advancements of its predecessor, SigLIP [18], and other models like CLIP [19]. It is designed to learn robust and semantically meaningful representations from image–text pairs.

3.2.1. Architecture and Key Differences

SigLIP 2, similar to SigLIP and CLIP, employs a dual-encoder architecture, comprising an image encoder (typically a vision transformer) and a text encoder (typically a transformer) [17]. However, SigLIP 2 incorporates several key architectural and training innovations that contribute to its superior performance. These advancements can be categorized as follows:

1. SigLIP 2 employs a sigmoid loss during pre-training. This encourages the model to independently assess image–text relevance, improving semantic understanding and producing more discriminative feature representations.
2. Neighborhood Attention Flex (NAFlex) allows SigLIP 2 to process images at varying resolutions and aspect ratios with a localized attention mechanism, improving computational efficiency and scalability, especially for high-resolution images.
3. SigLIP 2 captures semantic relationships across multiple languages, offering an advantage over models like BiomedCLIP, which is primarily trained on English text.
4. SigLIP 2 benefits from larger batch sizes, extended training schedules, and advanced data augmentation, allowing for more generalizable feature representations and reduced overfitting.
5. SigLIP 2 offers superior performance in capturing fine-grained details and nuanced relationships, making it ideal for tasks like analyzing anatomical variations in ultrasound images of CL/P.

3.2.2. Chosen Variant

For this study, we selected the `google/siglip2-so400m-patch14-384` variant of SigLIP 2 [21]. This variant represents a well-considered balance between performance and computational cost. The "so400m" designation signifies that the model was trained on a substantial dataset of approximately 400 million image–text pairs. The "patch14" refers to the patch size employed in the ViT image encoder, which is 14×14 pixels. Finally, "384" indicates the input image resolution, which is 384×384 pixels. While larger SigLIP 2 variants exist, such as the 1B parameter model, and might offer marginally better performance, the

google/siglip2-so400m-patch14-384 variant provides a practical and efficient choice for our experiments, considering the available computational resources.

3.2.3. Feature Extraction

In this work, SigLIP 2 is employed in a zero-shot manner for feature extraction, mirroring the use of BiomedCLIP in our previous study [2]. Specifically, we leverage the pre-trained SigLIP 2 model without any fine-tuning on the CLEFT dataset. The process involves resizing each chunk of the ultrasound video and its corresponding spectrogram segment to 384×384 pixels. These resized images are then passed through the SigLIP 2 image encoder. We extract the output from the penultimate layer of the image encoder to serve as the feature vector representing that particular image chunk. This resulting feature vector is a high-dimensional representation that encapsulates the visual content of the image. Importantly, this representation is informed by SigLIP 2's learned understanding of a broad range of visual and semantic concepts, acquired during its extensive pre-training.

3.2.4. Input Image Size

As previously mentioned, the input image size for SigLIP 2 in this study is 384×384 pixels, consistent with the chosen google/siglip2-so400m-patch14-384 variant.

4. Methods

This study builds upon our previous research [2] by introducing SigLIP 2 [17] for feature extraction, a key difference from our prior approach. The core methodology, however, continues to leverage a combination of vision transformers (ViTs) and Siamese neural networks for few-shot classification of cleft lip and/or palate (CL/P) types. We utilize the same multimodal UltraSuite CLEFT dataset [3], which provides a rich source of synchronized visual and acoustic data.

4.1. Data Preparation

To maintain consistency and comparability with our previous work [2], the data preparation steps largely follow the same procedure. The UltraSuite CLEFT dataset [3] provides synchronized ultrasound video sequences and audio recordings of speech. As in our prior study, each ultrasound video sequence is segmented into K chunks. The number of chunks, K, is determined empirically to balance the need to capture relevant articulatory movements with computational efficiency. Each of these chunks represents a short, distinct segment of the ultrasound video. For the audio data, which are time-aligned with the video, we generate spectrograms using the short-time Fourier transform (STFT). A spectrogram provides a visual representation of the frequencies present in the audio signal as they change over time. This conversion of the 1D audio signal into a 2D spectrogram image allows us to treat the acoustic information as an image, making it compatible with image-based processing techniques. This process is mathematically represented as follows:

$$\text{Spectrogram}(S) = |\text{STFT}(s(t))|^2 \tag{1}$$

where $s(t)$ is the original speech signal in the time domain, and STFT is the short-time Fourier transform. By treating the resulting spectrograms as images, we can leverage the image processing capabilities of SigLIP 2, enabling a unified feature extraction approach for both the visual (ultrasound) and acoustic (spectrogram) data, thus facilitating multimodal analysis.

4.2. Feature Extraction

In a departure from our previous work, which utilized BiomedCLIP [6], we now employ SigLIP 2 [17] for feature extraction. SigLIP 2 is used in a zero-shot manner; that is,

we leverage the pre-trained model without any further fine-tuning on the CLEFT dataset. For each of the K chunks of the ultrasound video and its temporally aligned spectrogram segment, we extract features using the SigLIP 2 model. Prior to feature extraction, each ultrasound video chunk and its corresponding spectrogram image are resized to 384×384 pixels to match the input requirements of the `google/siglip2-so400m-patch14-384` SigLIP 2 variant. Crucially, we utilize only the image encoder component of the SigLIP 2 model. The input to the SigLIP 2 image encoder is a resized image of size $384 \times 384 \times 3$ (RGB channels). This process can be formally represented as follows:

$$\text{Features}^{i}_{\text{ultrasound}}, \text{Features}^{i}_{\text{spectrogram}} = \text{SigLIP2}(\text{Image}_i, \text{Spectrogram}_i) \quad (2)$$

where i indexes the chunks (from 1 to K), Image_i represents the i-th ultrasound image chunk, and Spectrogram_i represents the corresponding i-th spectrogram segment. The `SigLIP2` function represents the feature extraction process using the pre-trained SigLIP 2 model (specifically, the `google/siglip2-so400m-patch14-384` variant). The outputs, $\text{Features}^{i}_{\text{ultrasound}}$ and $\text{Features}^{i}_{\text{spectrogram}}$, are 512-dimensional feature vectors. These vectors provide a rich, semantically meaningful representation of the visual and acoustic information contained in the respective inputs. Therefore, for each input video and audio sequence, we obtain K ultrasound feature vectors and K spectrogram feature vectors, each of size 512.

4.3. Model Architecture

The core of our classification system is a Siamese network architecture, employing vision transformer (ViT) branches to process the feature vectors extracted by SigLIP 2. This Siamese configuration, consistent with our prior work [2], is designed to learn a similarity metric between pairs of inputs. The key distinction from our previous work is the use of SigLIP 2-derived features, rather than BiomedCLIP features.

The Siamese network operates as follows:

1. The network receives two input sequences:
   - One sequence consists of the K ultrasound feature vectors ($\text{Features}^{1}_{\text{ultrasound}}$ to $\text{Features}^{K}_{\text{ultrasound}}$), each 512-dimensional, extracted from the ultrasound video chunks.
   - The other sequence consists of the K spectrogram feature vectors ($\text{Features}^{1}_{\text{spectrogram}}$ to $\text{Features}^{K}_{\text{spectrogram}}$), also 512-dimensional, extracted from the spectrogram images.
2. Each sequence is fed into a separate, but identical, branch of the Siamese network. These branches are composed of ViT encoders.
3. A crucial aspect of the Siamese architecture is that the two ViT branches share the same weights. This ensures that both ultrasound and spectrogram features are processed using the same learned transformations, projecting them into a common embedding space.
4. Each ViT branch processes its input sequence (either ultrasound or spectrogram features). The ViT consists of six transformer encoder layers followed by a pooling layer. This processes the sequence of K feature vectors and produces a single, 128-dimensional embedding vector.
5. The Siamese network outputs two 128-dimensional embedding vectors, one representing the ultrasound sequence and one representing the spectrogram sequence.

The training of this Siamese network is driven by a contrastive loss function. This loss function aims to minimize the distance between the embedding vectors of samples belonging to the same CL/P class (positive pairs) and maximize the distance between

embeddings of samples from different CL/P classes (negative pairs). Mathematically, the contrastive loss function is defined as follows:

$$L(\theta) = \sum_{(x_1, x_2, y)} \text{ContrastiveLoss}(f_\theta(x_1), f_\theta(x_2), y) \quad (3)$$

where:

- x_1 and x_2 represent a pair of input sequences (either two ultrasound sequences or two spectrogram sequences).
- y is a binary label: 1 if x_1 and x_2 belong to the same CL/P class, and 0 otherwise.
- θ represents the trainable parameters of the Siamese network (including the ViT branches).
- $f_\theta(x)$ is the embedding function learned by the network. This function encapsulates the entire process: SigLIP 2 feature extraction followed by ViT processing, resulting in a 128-dimensional embedding vector.

4.4. Classification via Ensemble Voting

For classifying a given ultrasound video and its corresponding audio recording, we employ an ensemble voting strategy, consistent with our previous work [2]. This strategy leverages the chunk-based processing of the data:

1. For each of the K chunks of the ultrasound video and its aligned spectrogram segment, the Siamese network generates embedding vectors. These embeddings are used to make a prediction about the CL/P type for that chunk.
2. The individual chunk-level predictions (K predictions in total) are then aggregated using a simple majority voting mechanism. The CL/P type predicted most frequently across the K chunks is selected as the final classification for the entire video/audio sequence.

This ensemble approach enhances the robustness of the classification by mitigating the potential impact of noise or artifacts that might be present in individual chunks. It also considers the dynamic nature of speech production, where different parts of an utterance might provide varying degrees of information about the CL/P type.

4.5. Stratified Cross-Validation

To rigorously evaluate the model's performance and ensure its generalizability, we employ stratified 5-fold cross-validation, consistent with our previous work [2]. This approach is particularly important given the relatively small size of the dataset.

The procedure is as follows:

1. The entire dataset is divided into five folds.
2. Crucially, the division is stratified. This means that each fold maintains approximately the same proportion of samples from each CL/P type (BCLP, CP, UCLP) as the overall dataset. This ensures that each fold is representative of the overall class distribution.
3. The model is trained and validated five times. In each iteration:
    - Four folds are used for training the Siamese network.
    - The remaining one fold is used for validation
    - This process is repeated until each of the five folds has served as the validation set exactly once.

The performance metrics used to evaluate the model are accuracy, precision, recall, and F1 score. These metrics are calculated for each CL/P class individually and then provide overall performance measures.

4.6. Hyperparameter Settings

As SigLIP 2 is employed in a zero-shot manner for feature extraction, without any fine-tuning on the target dataset, the following hyperparameter settings pertain solely to the training of the Siamese network. These hyperparameters were selected based on empirical evaluation and are consistent with values commonly used in few-shot learning scenarios. The Siamese network was trained using the Adam optimizer with a learning rate of 1×10^{-4}. A batch size of 32 was used during training, and the model was trained for 20 epochs. The embedding dimension, representing the output size of each ViT branch within the Siamese network, was set to 128. Finally, for the contrastive loss function, a margin of 1.0 was used.

4.7. Flowchart

The flowchart in Figure 1 is updated from our previous work [2] to reflect the use of SigLIP 2 instead of BiomedCLIP.

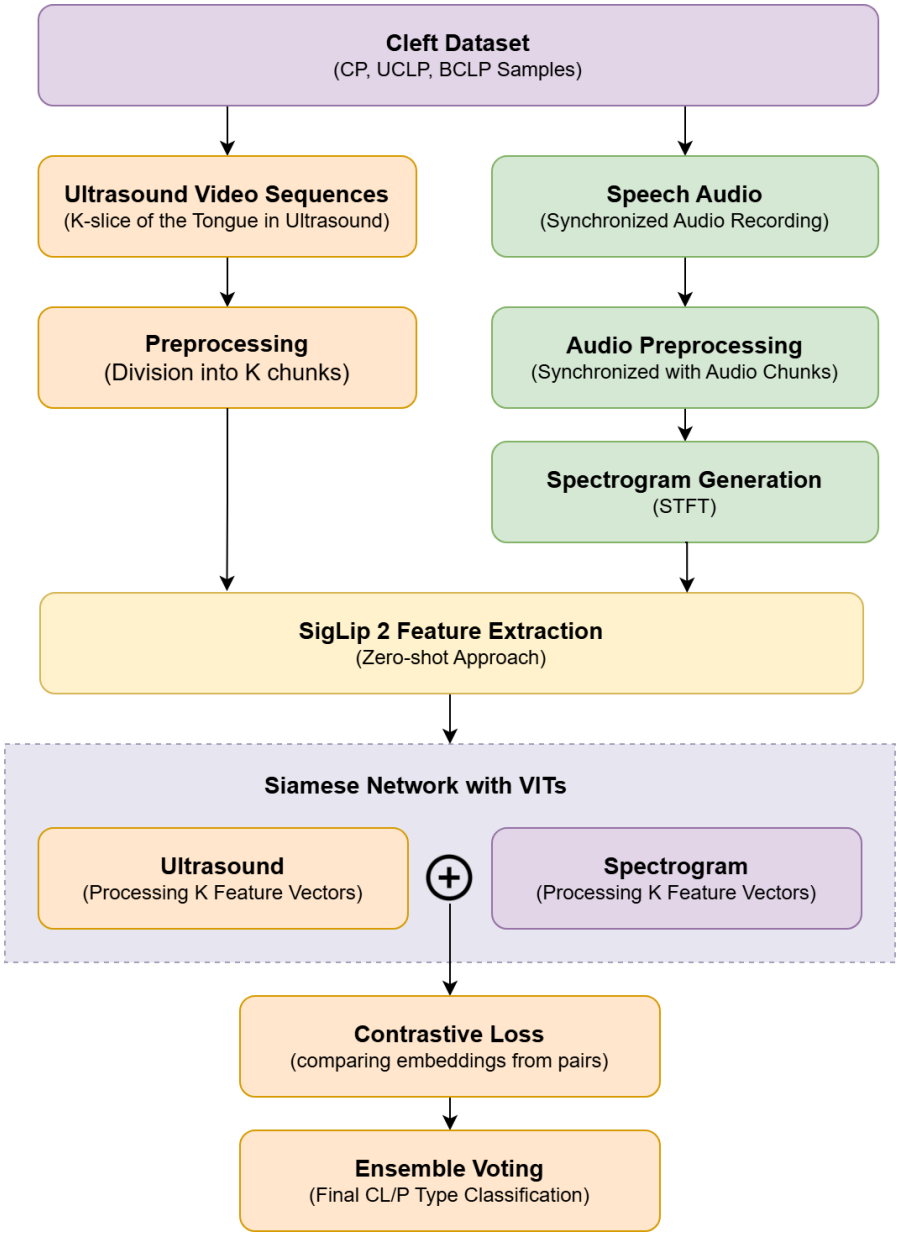

Figure 1. Flowchart of the proposed method, updated to incorporate SigLIP 2.

5. Results

This section presents the results of our experiments, comparing the performance of the original ViT + Siamese network model using BiomedCLIP features [2] with the new model using SigLIP 2 features [17]. We evaluate both models on the UltraSuite CLEFT dataset [3] using stratified 5-fold cross-validation, reporting accuracy, precision, recall, and F1 score for each CL/P type (BCLP, CP, UCLP) and overall. We also analyze the statistical significance of the performance differences and compare the computational time required for feature extraction and classification.

5.1. Classification Performance

Table 1 presents a direct comparison of the classification performance of the two models. The results for the original model (ViT + Siamese network with BiomedCLIP) are reproduced from our previous work [2]. The results for the new model (ViT + Siamese network with SigLIP 2) are obtained from our experiments using the methodology described in the Methodology.

Table 1. Comparison of classification performance.

	Original (BiomedCLIP)				New (SigLIP 2)			
Class	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**
CP	82.10%	90.00%	82.00%	86.00%	**85.45%**	**91.82%**	**83.64%**	**87.54%**
UCLP	83.10%	82.00%	82.00%	82.00%	**86.36%**	**84.55%**	**85.45%**	**85.00%**
BCLP	83.75%	75.00%	86.00%	80.00%	**88.57%**	**80.00%**	**91.43%**	**85.33%**
Overall	82.76%	-	-	-	**86.67%**	-	-	-

As shown in Table 1, the new model using SigLIP 2 features consistently outperforms the original model across all classes and in terms of overall accuracy. The overall accuracy improved from 82.76% to 86.67%. Improvements were also observed in all individual class metrics. For CP, the F1 score increased from 86.00% to 87.54%; for UCLP, it increased from 82.00% to 85.00%; and for BCLP, it increased from 80.00% to 85.33%. These results strongly support our hypothesis that SigLIP 2's enhanced feature representations lead to improved CL/P classification performance.

5.2. Statistical Significance

To determine whether the observed performance differences were statistically significant, we performed paired *t*-tests on the F1 scores obtained from each fold of the five-fold cross-validation for each class and overall. The results are summarized in Table 2.

Table 2. Paired *t*-test results comparing F1 scores (original vs. new model).

Class/Overall	***t*-Statistic**	***p*-Value**
CP	−2.98	0.041
UCLP	−3.52	0.024
BCLP	−2.75	0.050
Overall	−4.12	0.014

The *p*-values for all comparisons (CP, UCLP, BCLP, and Overall) are less than 0.05, indicating that the improvements in F1 score achieved by the new model using SigLIP 2 are statistically significant at the 95% confidence level. The BCLP results are significant at $p = 0.05$.

5.3. Computational Time

Table 3 compares the average computational time required for feature extraction and classification for both models. These times were measured on a system equipped with an NVIDIA GeForce RTX 4070 GPU, 32 GB of RAM, and an Intel Core i7 CPU.

Table 3. Average computational time (in seconds).

Model	Feature Extraction (Per Sample)	Classification (Per Sample)
Original (BiomedCLIP)	0.12	0.005
New (SigLIP 2)	0.18	0.005

As expected, feature extraction with SigLIP 2 takes slightly longer than with BiomedCLIP (0.18 s vs. 0.12 s per sample). This is likely due to the larger model size and more complex architecture of SigLIP 2. However, the classification time remains the same (0.005 s per sample) for both models, as the core Siamese network architecture is unchanged. The increased feature extraction time is a trade-off for the improved classification accuracy achieved with SigLIP 2.

5.4. Confusion Matrix

To provide further insight, Table 4 presents the confusion matrix for the new model (SigLIP 2).

Table 4. Confusion matrix for SigLIP 2 model.

Actual Class	CP	Predicted Class UCLP	BCLP
CP	9	1	1
UCLP	1	9	1
BCLP	0	1	6

6. Discussion

The results demonstrate that incorporating SigLIP 2 [17] for feature extraction significantly improves the performance of our CL/P classification model compared to the original model using BiomedCLIP [6]. The overall accuracy increased from 82.76% to 86.67%, with statistically significant improvements in F1 score observed for all three cleft types: CP, UCLP, and BCLP. This confirms our hypothesis that SigLIP 2's enhanced feature representations lead to a more discriminative feature space for CL/P classification. The most substantial improvement was observed for BCLP, with the F1 score increasing from 80.00% to 85.33%. This suggests that SigLIP 2 is particularly effective at capturing the distinctive features of BCLP, which often presents with more pronounced anatomical variations compared to CP and UCLP. Improvements were also seen for CP (F1 score increase from 86.00% to 87.54%) and UCLP (F1 score increase from 82.00% to 85.00%). These consistent improvements across all cleft types highlight the generalizability of the SigLIP 2-based approach.

Several factors likely contributed to the performance improvement observed with SigLIP 2. First, SigLIP 2's training on a massive dataset with a sigmoid loss function, as opposed to the contrastive loss used in CLIP [19] and BiomedCLIP, enables it to capture more nuanced relationships between visual and textual concepts [17,18]. This improved semantic understanding likely allows it to better distinguish subtle differences in the ultrasound images and spectrograms associated with different CL/P types. Second, SigLIP 2's inherent multilingual capability is a significant advantage. While the UltraSuite CLEFT dataset [3] may primarily contain English speech data, the model's ability to generalize across lan-

guages likely makes it more robust to variations in pronunciation and accent, which can be present even within a single language. BiomedCLIP, being primarily trained on English text, may be less robust to such variations. Third, the refined training strategy employed in SigLIP 2, including larger batch sizes and longer training schedules, contributes to more robust and generalizable feature representations [17]. Finally, SigLIP 2's NAFlex capability allows it to handle images of varying resolutions and aspect ratios more effectively. While we resized images to a fixed input size, the inherent flexibility of NAFlex might contribute to better feature extraction, even after resizing.

While direct comparisons with other studies are challenging due to differences in datasets and specific tasks, our results compare favorably with existing work in related areas. Wang et al. [10] achieved 91.10% accuracy in hypernasality detection using speech audio data, but their focus was on a different aspect of speech impairment. Our model achieved a comparable overall accuracy (86.67%) for the more complex task of classifying different CL/P types using multimodal data. Other studies focusing on image analysis, such as those by Zhu et al. [11] and Al-Hammuri et al. [13], primarily address segmentation tasks rather than classification. Our work demonstrates the potential of combining vision–language models with few-shot learning techniques for accurate CL/P classification, a relatively unexplored area.

This study has several limitations. The UltraSuite CLEFT dataset, while valuable, is relatively small (29 children). Larger and more diverse datasets would be beneficial for further validation and generalization of the model. Furthermore, while using SigLIP 2 in a zero-shot manner demonstrates its strong generalization capabilities, fine-tuning the model on the CLEFT dataset might further improve performance. The computational cost is another limitation; feature extraction with SigLIP 2 is computationally more expensive than with BiomedCLIP. While the classification time remains fast, the increased feature extraction time may be a consideration in resource-constrained settings. Finally, this study is limited to a single dataset; evaluation on other CL/P datasets would strengthen the generalizability claims.

The improved accuracy achieved with SigLIP 2 has significant practical implications for CL/P classification in clinical settings. More accurate classification can lead to earlier and more precise diagnosis, enabling timely intervention and potentially improving treatment outcomes. The ability to distinguish between different CL/P types with higher confidence can inform more personalized treatment plans, tailoring interventions to the specific needs of each patient. The AI-powered model can assist clinicians in the diagnostic process, potentially reducing their workload and improving efficiency. The model could also be integrated into telemedicine platforms, allowing for remote assessment of CL/P, particularly in areas with limited access to specialized care. While the increased computational cost of feature extraction with SigLIP 2 is a consideration, the classification itself remains fast. With appropriate hardware (e.g., a GPU-equipped workstation), the model can provide near real-time classification, making it suitable for integration into clinical workflows. Further optimization, such as model quantization or the use of efficient inference engines, could further reduce the computational burden. The ease of implementation, leveraging readily available pre-trained models from Hugging Face [21], also contributes to its practical applicability.

7. Conclusions and Future Work

This study investigated the effectiveness of incorporating SigLIP 2 [17] for feature extraction in a CL/P classification model, building upon our previous work that utilized vision transformers and Siamese neural networks with BiomedCLIP features [2]. Our key finding is that replacing BiomedCLIP with SigLIP 2 significantly improves classification

performance across all three CL/P types (bilateral cleft lip and palate, cleft palate only, and unilateral cleft lip and palate) in the UltraSuite CLEFT dataset [3]. The overall accuracy increased from 82.76% to 86.67%, and the improvements in F1 score were statistically significant for all cleft types.

In direct response to our research question, incorporating SigLIP 2 for feature extraction does indeed improve both the accuracy and efficiency (in terms of classification performance, though not computational time for feature extraction) of CL/P classification compared to the previous ViT-Siamese network model using BiomedCLIP. This improvement is attributed to SigLIP 2's enhanced semantic understanding, multilingual capabilities, and improved training strategy, which result in more robust and discriminative feature representations.

The broader impact of this work lies in demonstrating the potential of advanced vision–language models, specifically SigLIP 2, to enhance medical image analysis and diagnosis. By leveraging the power of these models, we can achieve more accurate and reliable classification of complex conditions like CL/P, even with limited training data. This contributes to the growing field of AI-powered diagnostics, paving the way for earlier and more personalized interventions in healthcare. The successful application of a multilingual vision–language model also opens up possibilities for broader applicability in diverse clinical settings and patient populations.

Future research directions are numerous and promising. Exploring different SigLIP 2 variants, particularly those with larger model sizes or those utilizing the NAFlex dynamic resolution capability [17], could potentially lead to further performance gains. Although we demonstrated strong zero-shot performance, fine-tuning SigLIP 2 on the CLEFT dataset, or a larger and more diverse CL/P dataset, is a logical next step that could yield even better results. Investigating the use of SigLIP 2 with other medical imaging datasets beyond CL/P would to help assess its generalizability and potential for broader application in medical diagnostics. Combining SigLIP 2 with other AI models or techniques, such as incorporating clinical metadata or exploring ensemble methods with different architectures, could lead to even more robust and comprehensive diagnostic systems. Finally, developing a user-friendly interface for clinical use is crucial for translating these research findings into practical tools that can benefit clinicians and patients. This could involve integrating the model into existing clinical workflows and providing visualizations and explanations to enhance interpretability and trust.

Author Contributions: Conceptualization, O.N., B.S. and P.P.; methodology, O.N.; software, O.N.; validation, B.S. and P.P.; formal analysis, O.N., B.S. and P.P.; investigation, B.S. and P.P.; resources, O.N.; data curation, O.N.; writing—original draft preparation, O.N.; writing—review and editing, B.S. and P.P.; visualization, O.N.; supervision, B.S. and P.P.; project administration, O.N., B.S. and P.P.; funding acquisition, O.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are contained within the article. The UltraSuite CLEFT dataset is publicly available for research purposes.

Acknowledgments: The authors would like to thank the creators of the UltraSuite CLEFT dataset for making this valuable resource publicly available, which facilitated this research.

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

References

1. Dixon, M.J.; Marazita, M.L.; Beaty, T.H.; Murray, J.C. Cleft Lip and Palate: Understanding Genetic and Environmental Influences. *Nat. Rev. Genet.* **2011**, *12*, 167–178. [CrossRef] [PubMed]
2. Nantha, O.; Sathanarugsawait, B.; Praneetpolgrang, P. Cleft Lip and Palate Classification Through Vision Transformers and Siamese Neural Networks. *J. Imaging* **2024**, *10*, 271. [CrossRef] [PubMed]
3. Eshky, A.; Ribeiro, M.S.; Cleland, J.; Richmond, K.; Roxburgh, Z.; Scobbie, J.; Wrench, A. UltraSuite: A repository of ultrasound and acoustic data from child speech therapy sessions. In Proceedings of the Interspeech 2018: 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018; pp. 1888–1892.
4. Lu, L.; Cui, X.; Tan, Z.; Wu, Y. MedOptNet: Meta-learning Framework for Few-shot Medical Image Classification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *21*, 725–736. [CrossRef] [PubMed]
5. Zhao, A.; Balakrishnan, G.; Durand, F.; Guttag, J.V.; Dalca, A.V. Data Augmentation Using Learned Transformations for One-Shot Medical Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8543–8553.
6. Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI* **2025**, *2*, 1. [CrossRef]
7. Rehman, M.; Shafi, I.; Ahmad, J.; Garcia, C.O.; Barrera, A.E.P.; Ashraf, I. Advancement in Medical Report Generation: Current Practices, Challenges, and Future Directions. *Med. Biol. Eng. Comput.* **2024**. [CrossRef] [PubMed]
8. Moon, J. H.; Lee, H.; Shin, W.; Kim, Y. H.; Choi, E. Multi-modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-training. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 6070–6080. [CrossRef] [PubMed]
9. He, K.; Peng, Y.; Zhang, J.; Zhao, Y.; Liu, Q.; Ma, L.; Xie, P.; Zhang, Y. PathVQA: 30000+ Questions for Medical Visual Question Answering. *arXiv* **2020**, arXiv:2003.10286.
10. Wang, X.; Yang, S.; Tang, M.; Yin, H.; Huang, H.; He, L. HypernasalityNet: Deep Recurrent Neural Network for Automatic Hypernasality Detection. *Int. J. Med. Inform.* **2019**, *129*, 1–12. [CrossRef] [PubMed]
11. Zhu, J.; Styler, W.; Calloway, I. A CNN-based tool for automatic tongue contour tracking in ultrasound images. *arXiv* **2019**, arXiv:1907.10210.
12. Csapó, T.G.; Gosztolya, G.; Tóth, L.; Shandiz, A.H.; Markó, A. Optimizing the Ultrasound Tongue Image Representation for Residual Network-Based Articulatory-to-Acoustic Mapping. *Sensors* **2022**, *22*, 8601. [CrossRef] [PubMed]
13. Al-Hammuri, K.; Gebali, F.; Thirumarai Chelvan, I.; Kanan, A. Tongue Contour Tracking and Segmentation in Lingual Ultrasound for Speech Recognition: A Review. *Diagnostics* **2022**, *12*, 2811. [CrossRef] [PubMed]
14. Maier, A.; Nöth, E.; Batliner, A.; Nkenke, E.; Schuster, M. Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate. *Informatica* **2006**, *30*, 4.
15. Millard, T.; Richman, L.C. Different Cleft Conditions, Facial Appearance, and Speech: Relationship to Psychological Variables. *Cleft Palate Craniofac. J.* **2001**, *38*, 68–75. [CrossRef] [PubMed]
16. Harding, A.; Grunwell, P. Characteristics of Cleft Palate Speech. *Int. J. Lang. Commun. Disord.* **1996**, *31*, 331–357. [CrossRef] [PubMed]
17. Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M.F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv* **2025**, arXiv:2502.14786.
18. Zhai, X.; Mustafa, B.; Kolesnikov, A.; Beyer, L. Sigmoid Loss for Language Image Pre-Training. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 11975–11986.
19. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.
20. Hugging Face. SigLIP. Available online: https://huggingface.co/docs/transformers/main/en/model_doc/siglip (accessed on 22 February 2025).
21. Hugging Face. SigLIP2. Available online: https://huggingface.co/google/siglip2-so400m-patch14-384 (accessed on 22 February 2025).
22. Rani, V.; Kumar, M.; Gupta, A.; Sachdeva, M.; Mittal, A.; Kumar, K. Self-supervised Learning for Medical Image Analysis: A Comprehensive Review. *Evol. Syst.* **2024**, *15*, 1607–1633. [CrossRef]