A Comparative Analysis of Traditional and Contemporary Visual Features for Computational Annotation of Irish Sign Language

Sarmad Khan[*†], **Simon D. McLoughlin**[†], **Irene Murtagh**[*†]
[*]ADAPT: AI-Driven Digital Content Technology, [†]Technological University Dublin
Dublin, Ireland
sarmad.khan@adaptcentre.ie, simon.d.mcloughlin@tudublin.ie, irene.murtagh@adaptcentre.ie

Abstract

Automatic annotation of sign language data is critical for advancing linguistic research and developing sign language technologies, yet it remains a major bottleneck due to the inherently motion-based and multi-modal nature of signing. Irish Sign Language, like many sign languages, presents challenges for computational annotation and sign language processing due to limited annotated corpora and the inherent difficulty of reliably annotating movement, trajectories, and coarticulation across manual and non-manual articulators. This paper presents an automated computational framework for gloss-level annotation support in Irish Sign Language, designed to assist scalable corpus annotation by learning motion-related cues directly from sign language videos. Using ELAN-aligned segments from the Signs of Ireland Corpus, we compare contemporary self-supervised visual representations with traditional pose-based features derived from explicit skeletal tracking, evaluating three feature configurations: DINOv2, MediaPipe, and multi-modal fusion. Our results show that self-supervised visual embeddings achieve the highest average accuracy 86.12 %, outperforming both multi-modal fusion 84.28 % and pose-based representations 76.74 %. This indicates that recent visual models can implicitly encode linguistically relevant motion information, including articulator movement and transitional dynamics, reducing the need for explicit landmark extraction in practical annotation pipelines. Overall, this work provides empirical guidance and a deployable computational framework to support computational annotation and enrichment of sign language corpora.

Keywords: Irish Sign Language, Automated Computational Annotation, Self-Supervised Learning, Sign Language Corpora

1. Introduction

The automatic annotation of Sign Language has the potential to advance corpus-based linguistic research, support the development of sustainable language resources, and facilitate sign language technologies that address communication barriers (Bragg et al., 2019). Sign languages are expressed through continuous movement: lexical and grammatical meaning is realised through coordinated movement trajectories of the hands, body, and face, unfolding continuously over time. Capturing this motion in a form that is both linguistically meaningful and computationally tractable remains a central challenge for the creation and analysis of sign language resources.

While substantial progress has been made for sign languages such as American Sign Language (ASL) and British Sign Language (BSL), Irish Sign Language (ISL) remains under-resourced with respect to annotated and data-driven processing resources (Khan et al., 2025). ISL is officially recognised as a native language of Ireland under the Irish Sign Language Act 2017. It serves as the first or preferred language for approximately 5,000 Deaf people in Ireland, with an estimated 40,000 people using ISL more broadly (Irish Deaf Society, 2024). Limited annotated resources, together with ISL's distinct grammar, gender-based linguistic variation, and multi-modal articulation, constrain progress in computational modelling of the language (Murtagh, 2021). Due to the visual-gestural nature of Sign Language, expert manual annotation is labour-intensive and difficult to scale, particularly when identifying gloss boundaries within continuous, coarticulated motion. Recent advances in artificial intelligence offer new support for sign language annotation by learning visual and temporal representations directly from video (Lugaresi et al., 2019; Caron et al., 2021). Pose-based methods explicitly model articulator movement via skeletal landmarks, whereas contemporary self-supervised vision models learn visual representations without manual feature design. For sign languages such as ISL, it remains unclear which visual feature extraction approaches best support scalable gloss-level annotation. From an annotation perspective, the central challenge is not only achieving high recognition accuracy, but enabling reliable computational annotation processes that reduce annotator workload across continuous, coarticulated signing.

This paper presents an automated computational annotation support pipeline for ISL that compares traditional pose-based representations with contemporary self-supervised visual features, as well as their multimodal fusion. We focus on evaluating how different representations encode motion across gloss-level sign segments and how this impacts their suitability for annotation-oriented workflows. Using ELAN-aligned data from an ISL cor-

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 239–247

pus, we conduct a comprehensive ablation study across 24 model configurations (8 classifiers $\times$ 3 feature representations) to assess representation robustness, efficiency, and scalability.

The main contributions of this paper are twofold:

- A comparative benchmark of traditional pose-based features (MediaPipe), contemporary self-supervised visual features (DINOv2), and their multimodal fusion for ISL gloss-level annotation support, highlighting the role of implicit motion representation.

- An empirical insight relevant to sign language corpora: self-supervised visual embeddings outperform multimodal fusion in both accuracy and computational efficiency, suggesting that implicit motion cues may suffice for scalable annotation support without explicit skeletal modelling.

The remainder of this paper is structured as follows: Section 2 reviews related work on sign language corpora and motion-based representations, Section 3 describes the dataset and preprocessing pipeline, Section 4 details the methodology, Section 5 presents experimental results, Section 6 discusses implications for sign language annotation and resource development, and Section 7 concludes the paper.

2. Related Work

Research on computational processing of sign languages has expanded rapidly in recent years, driven by advances in computer vision and machine learning aimed at improving accessibility and supporting linguistic analysis. Early work in sign language recognition primarily relied on convolutional neural networks and handcrafted motion descriptors to model visual gestures. Studies on Irish Sign Language (ISL) demonstrated that motion-based representations, such as Motion History Images and CNNs, can effectively capture hand shape and movement for isolated recognition tasks (Khan et al., 2024, 2025). While effective for constrained settings such as fingerspelling, these approaches are limited in their ability to represent the complex spatio-temporal dynamics required for gloss-level annotation in continuous signing.

Subsequent research introduced pose estimation to explicitly model articulator movement using skeletal landmarks. Frameworks based on MediaPipe and similar toolkits enabled the extraction of hand, body, and facial keypoints, which were combined with CNNs, recurrent networks, or attention mechanisms to improve recognition performance (Shukla and Gupta, 2025; Siju and Selvam, 2024). These approaches demonstrated that explicit modelling of articulator trajectories can support sign recognition, particularly in real-time or resource-constrained environments. However, pose-based pipelines introduce additional complexity and remain sensitive to occlusion, viewpoint variation, and tracking errors, which are especially problematic in large-scale corpus annotation. More recent work has shifted toward transformer-based architectures and self-supervised learning, which learn visual representations directly from raw video without manual feature specification. Vision transformers and masked-feature learning have been shown to capture rich spatial relationships and short-range temporal context through attention mechanisms, enabling more flexible modelling of sign language motion (Ravi Kiran et al., 2025; Caron et al., 2021). These representations have proven effective for both isolated and continuous sign language tasks and reduce reliance on explicit skeletal tracking. Parallel to advances in recognition models, several studies have highlighted the challenges of manual sign language annotation. Identifying sign boundaries within continuous motion, dealing with coarticulation, and integrating manual and non-manual features make annotation slow, subjective, and difficult to scale (Tian et al., 2024). Annotation is extremely time-intensive, with detailed linguistic annotation often requiring hours per minute of video, limiting the size and diversity of available corpora. Automated and semi-automated annotation tools have therefore become an important research focus, particularly for low-resource sign languages. Similarly, Low et al. (Low et al., 2025) propose a segment-aware visual tokenisation framework for gloss-free sign language translation, achieving computational efficiency through semantically informed temporal reduction. This work demonstrates that segment-level visual representations can reduce sequence length by 50% while maintaining translation quality.

Despite these developments, relatively little work has systematically examined which visual representations are most suitable for scalable, annotation-oriented processing of low-resource sign languages such as Irish Sign Language. Much of the existing research has focused on resource-rich sign languages, particularly American Sign Language (ASL) and British Sign Language (BSL), where large annotated datasets enable the use of complex models and extensive training regimes. In contrast, sign languages like ISL face additional constraints related to limited data availability, uneven distribution of gloss categories, and restricted annotation capacity. Many existing approaches focus either on pose-based or appearance-based features in isolation, and naive multi-modal fusion strategies remain underexplored in annotation contexts, especially for low-resource settings. This work addresses this gap by evaluating self-supervised visual representations, pose-based features, and their fusion within a unified computational annotation pipeline,

with an emphasis on motion representation, computational efficiency, and practical deployment for sign language corpus development.

3. Dataset and Preprocessing

3.1. Corpus Description

This study uses the **Signs of Ireland Corpus**, a linguistically annotated collection of Irish Sign Language (ISL) recordings developed at Trinity College Dublin (Leeson, 2004). The data comprises approximately 10 hours of naturalistic narrative signing collected in 2004 as part of a corpus development project involving members of the Irish Deaf community and researchers at Trinity College Dublin (Leeson, 2004). The corpus includes 40 Deaf native or early learners of ISL (24 female, 16 male), spanning four age groups (18–30, 30–45, 45–60, 65+) and five regions across Ireland (Dublin, Galway, Wexford, Waterford, Cork). Recordings consist primarily of narrative stories and retellings, including a subset of the Frog Story dataset, with time-aligned gloss annotations provided in ELAN XML format.

3.2. ELAN Parsing and Gloss Alignment

Each video in the Signs of Ireland Corpus is accompanied by an ELAN XML file containing time-aligned annotations for every gloss instance (Wittenburg et al., 2006). These annotations define precise temporal boundaries for the start and end of each sign, enabling frame-level segmentation and feature extraction. Formally, each gloss annotation is represented as:

$$A_i = \{ID_i, t_{\text{start}}, t_{\text{end}}, g_i\} \tag{1}$$

where ID_i is the unique annotation identifier, $t_{\text{start}}, t_{\text{end}}$ is the temporal boundaries of the gloss (in milliseconds) and g_i lexical gloss label.

To ensure consistency between the video data and its annotations, each gloss was automatically matched to its corresponding video file using the `MEDIA_URL` field embedded in the ELAN XML metadata. This process established a one-to-one mapping between gloss annotations and their respective video segments. After automated alignment and validation, a total of 3,865 gloss instances were successfully mapped from ELAN annotations to video clips, forming the basis for subsequent frame extraction and feature computation.

3.3. Gloss Filtering and Curation

To ensure high-quality input for training and evaluation, a four-stage filtering pipeline was applied to the raw ELAN gloss annotations:

1. **Frequency Thresholding**: Only glosses with at least 20 distinct occurrences were retained to ensure sufficient representation and statistical reliability per class.
2. **Lexical Normalisation**: All gloss labels were converted to `UPPERCASE` and cleaned by removing non-linguistic characters (e.g., `!`, `?`, numeric suffixes). The final curated the vocabulary of 55 glosses consists entirely of established lexical forms (e.g., `MOTHER`, `WORK`, `DAY`), as compound and reduplication variants did not meet the minimum frequency threshold for inclusion.
3. **Linguistic Exclusion**: Non-lexical, ambiguous, or structurally inconsistent items were excluded to improve dataset coherence, as follows:
   - *Classifiers* (e.g., `CL-THING`, `CL-PERSON`): excluded as they represent productive, gradient forms rather than established lexical entries.
   - *Indexing signs* (e.g., `INDEX1`, `INDEX2`): removed as they are referential and context-dependent rather than lexically stable.
   - *Name signs and anonymised placeholders* (e.g., `_JOHN`, `*XYZ`): excluded as they are person-specific and not generalisable.
   - *Emblems and expressive gestures* (e.g., `THUMBS-UP`, `GESTURE`): excluded as they lack consistent phonological form across signers.
4. **Manual Review**: The final subset was manually reviewed to ensure:
   - Semantic and visual consistency across all occurrences.
   - Preference for established lexical glosses (standardised, dictionary-like entries) over productive glosses (context-specific, creative, or infrequent signs) (Sutton-Spence, 2007).

As a result, the dataset was reduced from an initial pool of 3,865 annotated gloss instances to a curated vocabulary of 55 distinct ISL gloss labels. While the full corpus contained annotations across hundreds of glosses, filtering by frequency (≥ 20) and linguistic consistency retained approximately 1,950 gloss instances mapped to these 55 target glosses. This curation prioritises lexical consistency and sufficient per-class representation, which is critical for stable learning in less-researched sign language settings. Alternative preprocessing strategies would likely influence model behaviour. For example, retaining lower-frequency glosses would increase vocabulary coverage but introduce greater class imbalance and variability, potentially reducing

classification reliability. Similarly, less strict filtering may introduce inconsistencies in gloss labels, affecting both training stability and evaluation outcomes. The final dataset, therefore reflects a deliberate trade-off between linguistic coverage and statistical robustness.

3.4. Dataset Statistics

All gloss instances were uniformly sampled to a fixed number of frames to ensure consistent representation across samples. For each gloss instance with duration $[t_{\text{start}}, t_{\text{end}}]$, we extract 20 frames using uniform temporal spacing:

$$t_i = t_{\text{start}} + i \cdot \frac{t_{\text{end}} - t_{\text{start}}}{19}, \quad i \in \{0, 1, \ldots, 19\} \quad (2)$$

This ensures consistent temporal coverage regardless of the original gloss duration, with frames indexed from $i = 0$ (start) to $i = 19$ (end). Table 1 summarises the key properties of the curated ISL dataset used in this study, including vocabulary size, sampling strategy, and annotation format.

Table 1: Summary of Curated Dataset Statistics

Metric	Description
Gloss Vocabulary	55 distinct ISL gloss labels
Gloss Instances	∼1,950 annotated gloss segments
Frame Resolution	448×448
Sampling Strategy	20 uniformly spaced frames per gloss instance
Data Split	80% training / 20% testing (stratified by gloss)
Annotation Format	ELAN XML with gloss-level temporal segmentation

3.5. Feature Extraction

To construct semantically rich representations for Irish Sign Language gloss classification, we extract two complementary feature modalities from each video frame. These feature representations are designed to support gloss-level computational annotation by capturing motion cues that distinguish lexical signs within continuous signing.

3.5.1. DINOv2 Visual Features

DINOv2 is a self-supervised vision model that learns transferable features via self-distillation (Oquab et al., 2023). For each input frame I, the image is resized to 448×448 pixels and normalised to the range $[-1, 1]$ using standard scaling. The processed frame is passed through the DINOv2 ViT-S/14 encoder, yielding two levels of representation:

- **CLS Token Vector** ($\mathbf{f}_{\text{cls}} \in R^{384}$): A global descriptor obtained from the special classification token at the transformer's output. It captures coarse-level context such as signer pose, location, and signing space.
- **Patch Tokens** ($\mathbf{P} \in R^{256\times384}$): Representations corresponding to 16×16 spatial patches of the input frame, enabling localised visual analysis.

To remove background noise and emphasise signer-specific regions, we apply **PCA-based foreground selection** over the patch tokens:

$$M_{\text{fg}} = \left\{ p_i \in P \;\middle|\; \frac{p_i - \min(p)}{\max(p) - \min(p)} < 0.4 \right\} \quad (3)$$

The selected foreground tokens are averaged to produce a local structure embedding:

$$\mathbf{f}_{\text{patch}} = \frac{1}{|M_{\text{fg}}|} \sum_{p \in M_{\text{fg}}} p \quad (4)$$

Finally, the visual feature vector is constructed by concatenating the global and local descriptors:

$$\mathbf{f}_{\text{visual}} = [\mathbf{f}_{\text{cls}}; \mathbf{f}_{\text{patch}}] \in R^{768} \quad (5)$$

3.5.2. MediaPipe Skeletal Features

To complement appearance features, we extract skeletal motion and pose information using MediaPipe Holistic (Grishchenko and Bazarevsky, 2020). For each frame, the following landmark sets are obtained:

- **33 Body Pose Landmarks:** Each with (x, y, z, v) coordinates, giving $33 \times 4 = 132$ features.
- **21 Hand Landmarks per Hand:** Both hands are represented using (x, y, z) for each keypoint, totaling $2 \times 21 \times 3 = 126$ features.

The full skeletal vector is then formed as:

$$\mathbf{f}_{\text{pose}} = [\text{pose}; \text{hands}] \in R^{258} \quad (6)$$

This pose vector captures articulator positions, joint geometry, and inter-limb relationships relevant to distinguishing ISL glosses.

3.6. Multi-modal Feature Fusion

To integrate complementary information from both modalities, we concatenate the DINOv2-based visual embeddings with the MediaPipe-based skeletal features:

$$\mathbf{f}_{\text{multi}} = [\mathbf{f}_{\text{visual}}; \mathbf{f}_{\text{pose}}] \in R^{1026} \quad (7)$$

This combined representation encodes both high-level semantic content (via DINOv2) and

fine-grained spatial articulations (via pose landmarks), enriching the feature space for classification. Table 2 provides a detailed summary of all extracted features, their dimensionalities, and semantic meanings:

Table 2: Extracted Feature Representations

Feature	Dim.	Description
DINOv2 CLS	384	Global visual embedding
DINOv2 Patch Mean	384	PCA-selected foreground patches
MediaPipe Body	132	33 body landmarks (x, y, z, v)
MediaPipe Hands	126	21 landmarks $\times$ 2 hands (x, y, z)
Visual (DINOv2)	**768**	CLS + patch features
Pose (MediaPipe)	**258**	Body + hand landmarks
Multi-modal	**1026**	Visual + pose features

4. Methodology

4.1. Experimental Design

We conduct a comparative ablation study to systematically identify the most effective feature representation for Irish Sign Language (ISL) gloss classification. The experimental framework evaluates three distinct feature configurations MediaPipe (traditional skeletal tracking), DINOv2 (contemporary vision features), and their multi-modal fusion across eight diverse classification algorithms spanning gradient boosting, kernel methods, neural networks, and probabilistic models.

This design enables rigorous comparison of traditional versus contemporary approaches while empirically testing whether multi-modal fusion yields complementary benefits over single-modality representations. Figure 1 illustrates the comprehensive methodological framework employed for ISL gloss classification, detailing each stage from data preprocessing to model evaluation.

4.2. Feature Configurations

We evaluate three feature representations that embody different paradigms for SL understanding (detailed extraction methodology provided in Section 3.5):

4.2.1. MediaPipe (Traditional Skeletal Features)

Explicit articulator tracking via Google MediaPipe Holistic, extracting 258-dimensional skeletal keypoint features comprising 33 pose landmarks (x, y, z, visibility) and 21 hand landmarks per hand (x, y, z). This traditional approach leverages geometric relationships between hand configurations, spatial positions, and body pose to represent signing gestures through explicit articulator trajectories.

4.2.2. DINOv2 (Contemporary Vision Features)

Holistic visual representation via Meta AI's DINOv2-ViT-S/14 self-supervised vision transformer, yielding 768-dimensional feature vectors. We extract the 384-dimensional CLS token capturing global image semantics and mean-pool the top 40% most salient patch features (384 dimensions), selected via PCA-based saliency analysis to isolate signing-relevant regions. This contemporary approach captures holistic visual patterns including motion blur, spatial context, hand-object interactions, and appearance cues without requiring explicit keypoint detection.

4.2.3. Multi-modal Fusion

Simple early fusion concatenates DINOv2 visual features (768D) with MediaPipe skeletal features (258D), yielding a (1026D) multi-modal representation. This configuration tests whether combining global visual patterns with explicit articulator geometry provides complementary discriminative information for robust ISL gloss annotation.

4.3. Classifier Diversity

We evaluate three feature representations across eight standard classifiers: XGBoost, CatBoost, Random Forest, Support Vector Machines, Logistic Regression, Multi-Layer Perceptron, K-Nearest Neighbors, and Naive Bayes spanning ensemble, kernel-based, neural, probabilistic, and instance-based learning paradigms to ensure a robust and model-agnostic comparison.

4.4. Training Protocol and Hyperparameter Optimisation

All classifiers are trained using an 80:20 stratified train–test split. Hyperparameters follow standard best practices with empirical tuning and early stopping to control overfitting. Table 3 summarises the training configurations adopted for each classifier.

4.5. Evaluation Metrics

We evaluate classification performance using a standard set of metrics, capturing both overall accuracy and class-level behaviour. These include accuracy, macro- and weighted F1-scores to account for class imbalance, as well as Top-K accuracy to reflect annotation-oriented use cases where multiple candidate glosses may be presented. Model performance is evaluated against expert-annotated ELAN gloss labels from the Signs of Ireland Corpus, which serve as the ground truth. Predictions

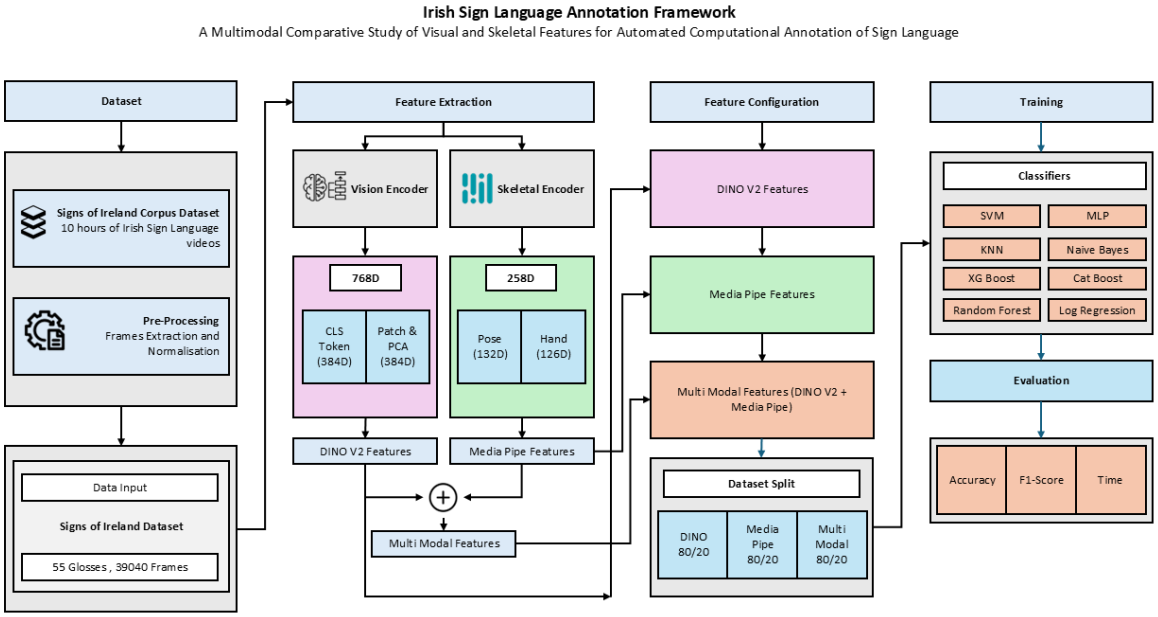

Figure 1: End-to-end computational pipeline for Irish Sign Language gloss annotation, illustrating dataset preparation, visual and skeletal feature extraction, multi-modal fusion, and model evaluation.

Table 3: Key Hyperparameter Configurations

Classifier	Configuration
XGBoost	1000 trees, depth 6, lr 0.15, early stopping
CatBoost	1000 iters, depth 6, lr 0.1, early stopping
Random Forest	300 trees, max depth 20, Gini split
SVM	RBF kernel, C=1.0
Logistic Regression	L2 regularisation, C=1.0, 1000 iters
MLP	Hidden layers (200,100), ReLU, reg $= 0.001$
KNN	k=7, Euclidean distance, weighted voting
Naive Bayes	Smoothing $= 10^{-8}$

Table 4: Evaluation Metrics

Metric	Description
Accuracy	Overall proportion of correct predictions
Top-K Accuracy	True label appears among the top K predicted labels
Macro F1 Score	Average F1 score across all classes (equal weighting)
Weighted F1 Score	F1 score adjusted for class imbalance
Log Loss	Measures prediction confidence and penalises errors
Per-Class F1	F1 score reported individually for each class

Table 5: Implementation Environment and Tools

Component	Tools and Libraries
Programming	Python 3.10+
Visual Features	PyTorch 2.0+, DINOv2 (via `torch.hub`)
Skeletal Features	MediaPipe Holistic v0.10.0
Hardware	GPU-accelerated (Google Colab Pro, CUDA-enabled)

are compared directly with these annotations using standard classification metrics.Table 4 summarises the metrics used.

4.6. Experimental Setup

All experiments were conducted in Python 3.10+ using modern machine learning and vision libraries. Table 5 outlines the key software and hardware setup. All experiments used fixed random seeds (`random_state=42`) for data splitting, initialisation, and validation. The complete experimental workflow is implemented, with structured JSON metadata stored for full reproducibility and traceability.

5. Results

5.1. Overall Performance

Across all classifiers, DINOv2 features achieve the highest average accuracy, outperforming both multi-modal fusion and MediaPipe-based representations. The strongest individual result is obtained

using XGBoost with DINOv2 features. Table 6 provides average accuracy across all classifiers.

Table 6: Average classification accuracy across all eight classifiers.

Feature Configuration	Avg. Accuracy (%)
DINOv2	**86.12**
MediaPipe	76.74
Multi-modal	84.28

5.2. Cross-Model Comparison

Several key observations emerge from this comprehensive comparison:

- **DINOv2 dominance:** DINOv2-based features achieve the highest accuracy for the majority of classifiers, with particularly strong performance for gradient boosting methods, where XGBoost reaches a peak accuracy of 95.91%.
- **Performance Variation**: Naive Bayes performs poorly across all configurations (13.97–25.32%), as its feature independence assumption is violated by the correlated DINOv2 embeddings and kinematically dependent MediaPipe landmarks, rendering it unsuitable for sign language feature representations.
- **Limited benefit of fusion:** Multi-modal fusion does not consistently outperform DINOv2 representations and, in several cases, results in reduced accuracy, suggesting redundancy or interference between visual and pose features.
- **Classifier sensitivity:** MediaPipe features perform competitively for some ensemble models but degrade substantially for linear and probabilistic classifiers, indicating limited suitability of explicit skeletal representations for certain learning paradigms.
- **Model robustness:** Contemporary self-supervised visual embeddings generalise more reliably across diverse classifier families than pose-based or fused representations, supporting their use in annotation-oriented workflows.

5.3. Computational Efficiency Analysis

An important consideration for practical deployment is the trade-off between accuracy and computational cost. Table 7 compares average accuracy and training time across all eight classifiers for each feature configuration.

- **DINOv2:** Provides the best balance between accuracy and computational cost across classifiers, making it well suited for scalable annotation workflows.

Table 7: Model-wise efficiency comparison: accuracy vs. training time

Model	DINOv2		MediaPipe		Multi-modal	
	Acc.	Time	Acc.	Time	Acc.	Time
XGBoost	95.91	272	95.12	163	95.61	3,881
CatBoost	95.84	299	95.17	134	94.74	4,646
Random Forest	95.43	13	94.83	7	85.07	20
MLP	95.26	217	91.09	117	94.74	15
KNN	95.35	0.1	86.74	0.03	92.57	42
Log. Reg	94.45	50	60.05	22	94.56	4
SVM	93.99	96	76.96	47	91.61	7
Naive Bayes	22.69	0.5	13.97	0.1	25.32	0.5
Avg.	86.12	119	76.74	61	84.28	1,077

- **Multi-modal:** Incurs substantially higher training cost without corresponding accuracy gains, indicating limited efficiency for practical deployment.
- **MediaPipe:** Trains fastest but yields lower accuracy, favouring speed-oriented scenarios over annotation quality.

5.4. Per-Class Analysis

To evaluate gloss-wise classification performance, we analyse per-class F1-scores from the best-performing model (XGBoost with DINOv2 features). Table 8 shows the 10 highest scoring and 10 lowest scoring glosses out of 55 total classes.

Table 8: Higher and lower scoring ISL glosses

Higher Scoring Glosses	F1	Lower Scoring Glosses	F1
DRINK	0.99	DAY	0.94
SMALL	0.99	TIME	0.94
MOTORBIKE	0.99	NOT	0.94
PLAY	0.99	LOOK	0.94
THREE	0.99	FUNNY	0.94
DEAF	0.99	FATHER	0.93
BOY	0.98	WHERE	0.93
SAY	0.98	MY	0.93
SISTER	0.98	FINISH	0.91
TWO	0.98	SAME	0.90

Overall, the model demonstrates strong and consistent performance across glosses, with 58.2% of classes achieving F1-scores of 97% or higher. Only a small proportion of glosses (10.9%) fall below 95% F1, indicating robustness even for less frequent or potentially ambiguous signs.

5.5. Confusion Matrix Analysis

Figure 2 presents the complete 55×55 confusion matrix for the best model, exhibiting strong diagonal concentration with 95.91 % correct classifications. The matrix demonstrates error distribution with predominantly small or zero off-diagonal elements, indicating that misclassifications are rare and not systematically biased.

6. Discussion and Conclusion

This work addresses automated computational annotation of sign language corpora, where the main

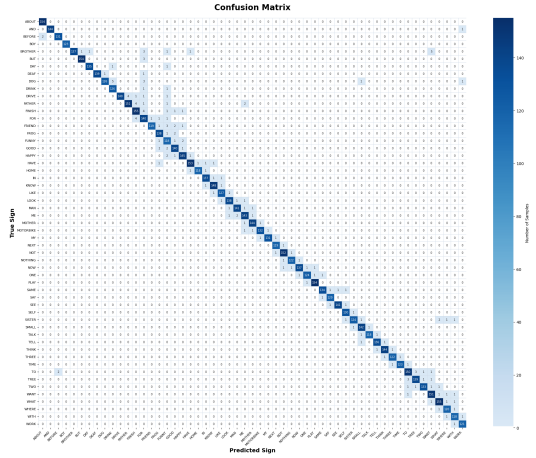

Figure 2: Confusion matrix for XGBoost with DINOv2 features

bottleneck is the cost and consistency of gloss-level labelling in continuous, motion-rich signing. For Irish Sign Language (ISL), this challenge is intensified by limited annotated data and the difficulty of identifying gloss boundaries under coarticulation. Our results show that self-supervised visual embeddings provide a robust baseline for annotation assistance on high-frequency glosses, consistently outperforming pose-based and multi-modal representations across classifiers. From a corpus development perspective, DINOv2 offers the most favourable accuracy–efficiency trade-off, achieving strong performance while avoiding the added complexity and computational cost of pose estimation and feature fusion. The limited benefit of naive multi-modal fusion further suggests that explicit skeletal landmarks may not be essential when contemporary visual models already capture linguistically relevant motion cues. These findings motivate practical recommendations for ISL annotation workflows: prioritising self-supervised visual representations, treating pose estimations as optional rather than essential, and employing efficient classifiers such as gradient boosting. In summary, this paper presents an automated pipeline for ISL gloss-level annotation and offers empirical insight into motion-aware visual representations that balance accuracy and computational efficiency. The approach supports faster and more consistent gloss suggestions for sign language corpus development. The set of 55 glosses reflects a focus on high-frequency, consistent signs to ensure reliable learning, which serves as a controlled benchmark. Future work will apply explainable artificial intelligence methods to better understand model decisions and refine the annotation framework.

Acknowledgements

This research was supported by the ADAPT Centre for AI-Driven Digital Content Technology at Technological University Dublin. The ADAPT Centre is funded by Research Ireland through the Research Centres Programme. The authors also thank the contributors and annotators involved in the creation of the Signs of Ireland Corpus.

Bibliographical References

Danielle Bragg, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic — simultaneous face, hand and pose prediction, on device. Google Research Blog.

Irish Deaf Society. 2024. Irish sign language. https://www.irishdeafsociety.ie/irish-sign-language/.

Hafiz Muhammad Sarmad Khan, Simon D. Mcloughlin, and Irene Murtagh. 2025. Comparative evaluation and utilization of convolutional neural network architectures for irish sign language recognition. *Journal of Artificial Intelligence*, 16(1).

Sarmad Khan, Irene Murtagh, and Simon D. McLoughlin. 2024. Investigating motion history images and convolutional neural networks for isolated irish sign language fingerspelling recognition. In *Proceedings of the LREC-COLING 2024 11th Workshop on Sign Languages*.

Lorraine Leeson. 2004. Signs of ireland corpus: a collection of irish sign language video data from 40 signers of ireland.

Jian He Low, Ozge Mercanoglu Sincan, and Richard Bowden. 2025. Sage: Segment-aware gloss-free encoding for token-efficient sign language translation. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 5011–5020.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Irene Murtagh. 2021. The nature of verbs in sign languages: A role and reference grammar account of irish sign language verbs. *TEANGA,*

the Journal of the Irish Association for Applied Linguistics, 11:67–99.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Gali Ravi Kiran, Varada Srinadh, V Ankitha, and S Surekha. 2025. Sign language translator using transformer model. *International Journal of Environmental Sciences*, 35(1):84–92.

Manish Shukla and Harsh Gupta. 2025. Advancing sign language interpretation with transfer learning and multimodal features. *Research Square Preprint*.

Ijas Siju and Prabu Selvam. 2024. A novel approach for lightweight sign language recognition leveraging google mediapipe and deep neural net. In *Proceedings of the 2024 International Conference on Signal Processing and Integrated Networks*, pages 746–751. IEEE.

Rachel Sutton-Spence. 2007. Signs in BSL – established or productive? Technical report, University of Bristol, Graduate School of Education.

Yingli Tian, Jianbo Su, Lan Ni, and Yuchun Fang. 2024. Editorial: Bridging the gap: Ai and sign language recognition– a path toward inclusive communication. *Frontiers in Computer Science*, 6:1203805.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research.