Jurnal Teknik Informatika (JUTIF)
Vol. 6, No. 1, February 2025, pp. 187-204

DOI: https://doi.org/10.52436/1.jutif.2025.6.1.4202
p-ISSN: 2723-3863
e-ISSN: 2723-3871

A STUDY OF WORLDWIDE PATTERNS IN ALPHABET SIGN LANGUAGE RECOGNITION USING CONVOLUTIONAL AND RECURRENT NEURAL NETWORKS

Aris Rakhmadi*[1,2], Anton Yudhana*[3], Sunardi*[3]

[1]Informatics, Universitas Ahmad Dahlan, Indonesia
[2]Informatics Engineering, Universitas Muhammadiyah Surakarta, Indonesia
[3]Electrical Engineering, Universitas Ahmad Dahlan, Indonesia
Email: [1]aris.rakhmadi@ums.ac.id, [2]eyudhana@ee.uad.ac.id, [3]sunardi@ee.uad.ac.id

(Article received: December 10, 2024; Revision: January 19, 2025; published: February 20, 2025)

Abstract

Sign Language Recognition (SLR) has become an essential area of research due to its potential to promote understanding between the deaf and hearing communities through communication. This paper provides an in-depth study of various methodologies and models employed in SLR, focusing on Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). We analyze their application to datasets from various sign languages, such as Arabic Sign Language (ArSL), American Sign Language (ASL), and British Sign Language (BSL), and explore how these models improve the recognition of dynamic, multi-dimensional hand gestures. This research not only advances the understanding of deep learning applications in sign language recognition but also addresses critical challenges in data processing and real-time applications, paving the way for inclusive technologies in informatics and human-computer interaction. Despite the progress in applying deep learning techniques to SLR, several challenges remain, particularly in dataset limitations, handling large vocabularies, and ensuring consistent performance across diverse environments and signers. The paper also investigates the broader applications of SLR, such as virtual reality, healthcare, education, and accessibility, and discusses the integration of SLR with human-computer interaction systems. Furthermore, it highlights current limitations in the field, such as difficulties with video data handling, the need for standard datasets, and issues related to training computational models. Finally, the paper outlines future research directions, including developing more robust SLR systems that can function effectively in uncontrolled environments, improving data collection methodologies, and creating real-time, user-friendly applications to assist the community of deaf and hard-of-hearing individuals.

Keywords: *Sign Language Recognition, Convolutional Neural Networks, Recurrent Neural Networks, Deep Learning.*

KAJIAN GLOBAL DALAM PENGENALAN POLA BAHASA ISYARAT ALFABET MENGGUNAKAN *CONVOLUTIONAL* DAN *RECURRENT NEURAL NETWORK*

Abstrak

Pengenalan Bahasa Isyarat (SLR) menjadi bidang penelitian yang sangat penting karena potensinya dalam memfasilitasi komunikasi dan meningkatkan pemahaman antara komunitas tuna rungu dan pendengaran. Artikel ini menyajikan kajian mendalam mengenai berbagai metodologi dan model yang diterapkan dalam SLR, dengan fokus pada penggunaan Convolutional Neural Networks (CNN) dan Recurrent Neural Networks (RNN). Penelitian mengkaji penerapan model-model tersebut pada dataset dari berbagai bahasa isyarat, seperti Bahasa Isyarat Arab (ArSL), Bahasa Isyarat Amerika (ASL), dan Bahasa Isyarat Inggris (BSL), serta mengeksplorasi bagaimana dapat meningkatkan pengenalan gerakan tangan yang dinamis dan multi-dimensional. Meskipun ada kemajuan dalam penerapan teknik *deep learning* untuk SLR, masih terdapat beberapa tantangan, terutama terkait keterbatasan dataset, pengelolaan kosa kata yang besar, dan memastikan kinerja yang konsisten di berbagai lingkungan serta pengguna isyarat. Artikel ini juga membahas aplikasi lebih luas dari SLR, seperti dalam realitas virtual, layanan kesehatan, pendidikan, dan aksesibilitas, serta menggali integrasi SLR dengan sistem interaksi manusia-komputer. Di samping itu, artikel ini menyoroti beberapa keterbatasan yang masih ada, seperti tantangan dalam penanganan data video, kebutuhan untuk dataset standar, serta masalah terkait pelatihan model komputasi. Artikel ini mengakhiri dengan menyarankan arah penelitian masa depan, termasuk pengembangan sistem SLR yang lebih kuat yang dapat beroperasi secara efektif dalam lingkungan yang tidak terkendali,

peningkatan metodologi pengumpulan data, dan pembuatan aplikasi real-time yang lebih ramah pengguna untuk membantu komunitas tuna rungu dan individu dengan gangguan pendengaran.

Kata kunci: *Sign Language Recognition, Convolutional Neural Networks, Recurrent Neural Networks, Deep Learning.*

1. INTRODUCTION

Sign languages are rich, visual languages utilized by deaf communities across the globe, each featuring its distinct grammatical structures and vocabulary. Unlike spoken languages, which rely on auditory cues, sign languages communicate meaning through hand gestures, facial expressions, and body movements. This visual modality allows for nuanced expression and conveys sometimes untranslatable ideas in spoken form. Each sign language is culturally specific, reflecting the unique social and linguistic contexts of the communities that use them, making them vital for fostering identity and communication among deaf individuals.

Between 138 and 300 unique sign languages are used globally, each characterized by its hand movements and facial expressions system. These languages possess distinct grammar rules and sentence structures, which can differ significantly from verbal languages. For instance, ASL has a different word order compared to English [1]. Additionally, sign languages typically rely on essential English word signals and do not include prepositional markers, complicating direct translation. Understanding these differences is critical for developing effective sign language translators that enable communication between sign language users and those unfamiliar with these languages [2].

Alphabet sign languages utilize specific hand shapes to represent individual letters of the alphabet, serving as a vital tool for bridging communication gaps between deaf and hearing individuals [3]. This fingerspelling system enables users to spell out names, places, or terms that may not have a designated sign, thus facilitating clear and accurate communication in various contexts, such as healthcare, education, and social interactions. By expressing words that lack established signs, alphabet sign languages enhance understanding and inclusion, allowing for more effective dialogue and fostering connections between diverse communities.

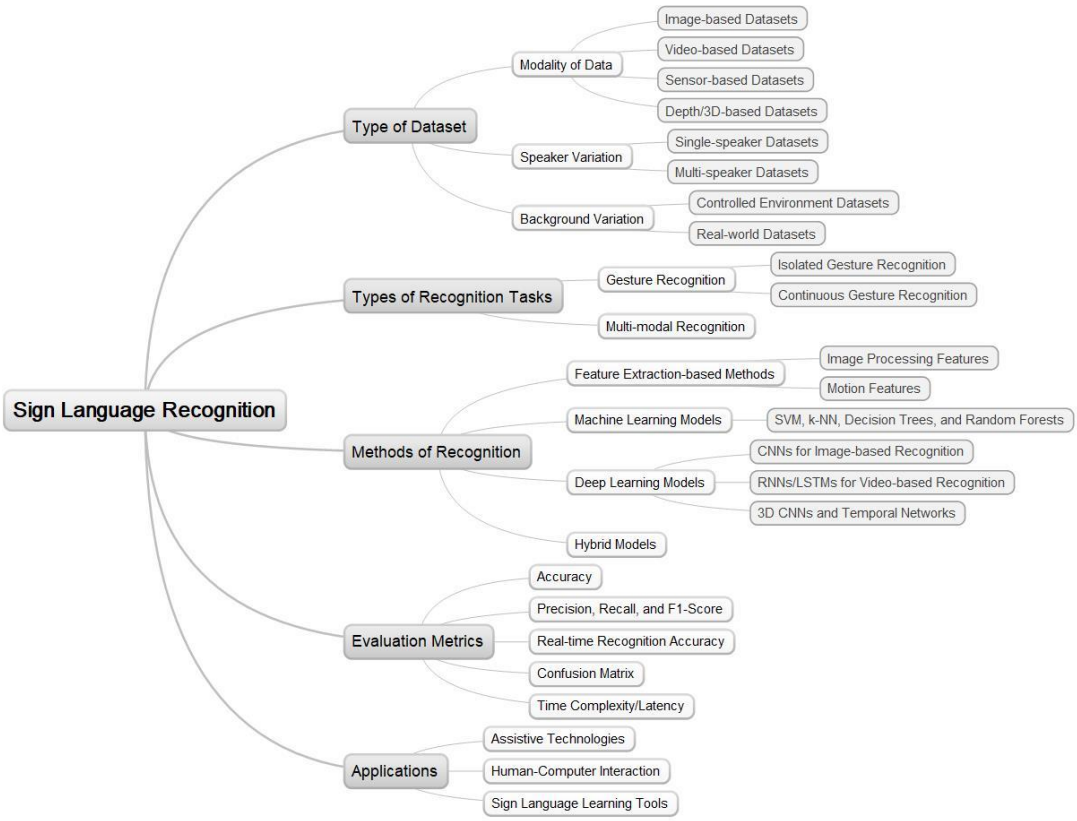


Figure 1. Taxonomy of Approaches and Methods in SLR

In SLR research, taxonomy refers to classifying methods, techniques, and challenges that define the field. Figure 1 shows the taxonomy diagram of SLR, which outlines the various categories that organize the diverse approaches and methods used in SLR research. The dataset type is a critical consideration in SLR, as it determines the form of data that will be used to train and test the recognition models. One significant distinction lies in the modality of the data, which refers to the type of input collected from the signer [4]. Image-based datasets focus on static images, capturing hand gestures or signs in a still frame [5]. These datasets are typically used for tasks involving the recognition of isolated signs, where the spatial configuration of the hand is the key feature. However, they cannot capture motion, making them less effective for recognizing dynamic or context-dependent signs. On the other hand, video-based datasets offer a dynamic perspective, capturing the motion of hand gestures over time [6]. This is especially crucial for recognizing continuous gestures or sentences, as it provides temporal information that allows the model to discern the progression of hand movements. Video datasets are more realistic but require more computational resources due to the large number of frames that need processing.

Another important modality is sensor-based datasets, where data is captured using wearable devices such as gloves equipped with sensors or accelerometers. These sensors can precisely capture the hand's position, shape, and movement, sometimes providing more detailed information than visual-based methods, especially regarding hand orientation and gesture accuracy [7]. Depth/3D-based datasets, such as those captured by depth cameras (e.g., Microsoft Kinect), use 3D sensors to capture the spatial position of the hand in three-dimensional space. These datasets are beneficial for recognizing complex gestures or subtle variations in hand position, as depth sensors can isolate hand movements from the background, overcoming issues like changes in lighting or cluttered backgrounds. Depth data is beneficial in handling occlusion and background interference challenges, providing a more robust solution for real-world SLR applications.

Speaker variation addresses the differences between individuals in their way of signing, significantly impacting the generalization of SLR systems. Datasets can be categorized as either single-speaker or multi-speaker. Single-speaker datasets focus on the performance of one individual signer, often capturing how that person produces a particular set of gestures or signs. These datasets can be helpful in controlled experiments or developing models targeting a specific signer [8], [9], [10]. However, their primary limitation is that the model trained on such data may struggle to generalize to other signers due to the lack of variability in signing styles, hand shapes, and gesture executions. This restricts the model's adaptability when faced with new or unseen signers, making it less effective in real-world applications where the system needs to work for a wide range of users.

In contrast, multi-speaker datasets capture a variety of signers, each with their unique sign style, physical characteristics (such as hand size and shape), and possibly even regional dialects of the same sign language. This broader representation is essential for training SLR systems that generalize across different users, making them more practical and robust in real-world scenarios. For example, American Sign Language (ASL) can vary between regions and communities, and multi-speaker datasets can help account for these variations [11]. Including multiple signers in the dataset ensures the model is not overfitted to a specific individual's characteristics, enhancing the system's ability to recognize signs from a diverse population [12].

Background variation refers to the environment in which the data is collected, and it plays a significant role in the performance of Sign Language Recognition systems [13]. Datasets can be categorized as controlled environment datasets or real-world datasets. In controlled environments, data collection is conducted in settings where external factors, such as lighting, background clutter, and other disturbances, are carefully controlled. This ensures that the focus remains solely on the gestures or signs being performed, which can help improve recognition accuracy. Controlled environments are ideal for early-stage model development, as they minimize variability, making it easier for models to learn the features of the gestures. However, such datasets often fail to capture the complexity and unpredictability of real-world situations, where signs may be performed under less-than-ideal conditions.

In real-world datasets, data is collected in natural, uncontrolled environments where various background conditions can influence the data. For instance, signs might be performed in public spaces with varying lighting and backgrounds or even during different times of the day. This introduces complexity that is difficult to replicate in a laboratory setting. While real-world datasets present challenges, they are crucial for developing robust SLR systems that perform well outside controlled settings [14]. These datasets help models learn to deal with variations in background, lighting, and other environmental factors, making them more adaptable to practical applications, such as in public spaces or at home [15], [16]. Real-world datasets are often more extensive and diverse, making them essential for ensuring that SLR systems are helpful in daily interactions.

In the field of SLR, recognition tasks define how the system processes and interprets the data collected from signers. These tasks vary in complexity based on whether the system identifies isolated gestures, continuous sign language sequences, or combines different input modalities for more accurate recognition [17]. The primary types of recognition tasks are gesture recognition and multi-modal recognition, each serving specific purposes and presenting unique challenges.

Gesture recognition is at the heart of SLR systems. The goal is to identify and interpret hand and body movements representing specific sign language signs or gestures. Gesture recognition tasks can be categorized into two subtypes: isolated gesture recognition and continuous gesture recognition [18].

Isolated gesture recognition involves the identification of a single, discrete gesture or sign, typically captured in isolation [19]. This task is more straightforward than continuous recognition because the system only needs to recognize one gesture at a time, with minimal temporal context. For instance, an SLR system might recognize a hand gesture such as the sign for "hello" or "thank you" in American Sign Language (ASL) without needing to account for the surrounding context or other signs. Isolated gesture recognition is helpful in applications where single signs must be translated or identified quickly, such as in a simple sign-to-text conversion system. However, it can be limited in handling more complex conversations, as it does not account for the flow of multiple signs or the interaction between gestures.

Continuous gesture recognition, on the other hand, deals with recognizing gestures in a constant flow, where multiple signs are performed in a sequence, often forming sentences or phrases [20]. Unlike isolated gesture recognition, this task requires the system to identify individual signs and understand their temporal relationships and contextual meaning within the sequence. Continuous gesture recognition is much more challenging due to the need to capture transitions between gestures, account for variations in speed, and manage the complexity of interpreting sentences. This is crucial for real-world applications, such as real-time sign language translation or communication systems, where the context and flow of signs must be accurately interpreted. Continuous recognition systems must process temporal features in the data, making them computationally more demanding than isolated gesture recognition.

Multi-modal recognition involves integrating data from multiple sources or modalities, such as combining visual data with sensor data (e.g., depth or motion sensors). This task is designed to improve the accuracy and robustness of SLR systems by leveraging complementary information from various input types [21].

In a multi-modal recognition system, the model typically combines data from visual sensors (such as cameras or depth sensors) and sensor-based inputs (such as accelerometers or IMUs) to capture a more comprehensive view of the signer's gestures. For example, a system might combine visual data to track hand shapes and movements with sensor data to capture fine-grained details like hand orientation or wrist rotation. The advantage of multi-modal recognition is that it allows the system to compensate for limitations inherent in any single modality. For instance, visual data might struggle in low-light conditions or with occlusions, but sensor data can provide additional, accurate information to ensure reliable recognition.

Multi-modal recognition is instrumental in complex or dynamic environments where a single input type is insufficient for accurate sign language recognition. By integrating different modalities, multi-modal systems can enhance performance, robustness, and accuracy, making them suitable for real-world applications, such as human-robot interaction, augmented reality, or assistive technologies for people with hearing impairments. However, the main challenge lies in fusing the data from different modalities to minimize noise and maximize the usefulness of each source [15]. Multi-modal systems are computationally more demanding and require advanced algorithms for data fusion. Still, their ability to handle diverse environments and improve recognition accuracy makes them highly valuable in practical SLR applications.

The recognition methods in Sign Language Recognition (SLR) involve various techniques for interpreting and classifying gestures based on the data collected from signers. These methods can be broadly categorized into feature extraction-based, machine learning, deep learning, and hybrid models. Depending on the type of data and the complexity of the recognition task, each of these methods uniquely improves the accuracy and efficiency of SLR systems [22].

Feature extraction-based methods are foundational in many traditional sign language recognition systems. These methods rely on extracting discriminative features from raw input data, which are subsequently used to classify or recognize specific gestures [23], [24]. The primary categories of features extracted are image processing features and motion features. Image processing features are derived from static images or video frames, where various image characteristics such as shape, texture, and contour are extracted to identify hand configurations. Techniques such as edge detection, corner detection, and keypoint extraction enable the system to map distinct hand landmarks,

such as the fingertips and joints, which serve as the basis for recognizing individual signs. This approach efficiently recognizes isolated hand gestures but fails to account for temporal information, critical for recognizing dynamic gestures or sign language sequences. Motion features, in contrast, are derived from the temporal progression of gestures, capturing the movement of hands or the body. Motion vectors, optical flow, and trajectory analysis are commonly used to quantify hand motion and provide a time-dependent representation of gestures [25]. These features are essential for continuous gesture recognition, where the system must discern individual signs and their transitions within a sequence of gestures. Thus, motion features enable the recognition of sign language that spans multiple frames or even complete sentences.

In parallel, machine learning methods have been widely applied to classify gestures based on the features extracted from raw data. Classical algorithms in machine learning, such as Support Vector Machines (SVM), K-nearest neighbors (K-NN), Decision Trees, and Random Forests, are used to develop models that can accurately predict the class of a given gesture [26], [27]. Support Vector Machines are particularly effective for high-dimensional feature spaces, as they construct an optimal hyperplane that maximally separates different gesture classes. K-Nearest Neighbors, by contrast, classifies gestures based on proximity to labeled instances in the feature space, providing a simple yet effective method for gesture recognition. Decision Trees create hierarchical models, partitioning the feature space through a series of decision nodes. At the same time, Random Forests improve upon this by combining multiple decision trees, thus reducing overfitting and improving generalization. These machine-learning techniques are particularly effective for tasks involving isolated gesture recognition. Still, they are often less efficient for continuous or complex sign language sequences, where temporal dependencies play a crucial role in understanding the meaning of signs.

In recent years, deep learning models have emerged as a dominant approach in sign language recognition due to their ability to learn hierarchical, abstract features from raw data directly. Convolutional Neural Networks (CNNs) are primarily used for image-based recognition, where their convolutional layers automatically learn spatial patterns within visual data, such as hand shapes, orientations, and configurations. CNNs excel at recognizing isolated hand gestures from images or video frames by learning a series of spatial filters that capture low- and high-level features. However, sign language recognition frequently involves dynamic gestures, necessitating the modeling of temporal dependencies. Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, address this challenge by processing sequential data and capturing long-range dependencies between gestures in a sequence. RNNs are designed to handle sequential inputs, such as frames in a video, making them well-suited for dynamic gesture recognition [28], [29]. LSTMs, an advanced variant of RNNs, are specifically designed to alleviate the vanishing gradient problem and can retain information over long sequences, making them ideal for continuous gesture recognition. 3D CNNs and temporal networks are employed for spatial and temporal information tasks. 3D CNNs extend traditional CNNs by adding dimension and time, allowing the network to learn spatial and temporal patterns simultaneously. These networks are particularly effective for recognizing dynamic gestures and sign language sequences where the motion of the hands is integral to the sign's meaning. Temporal networks, designed to model sequential dependencies more effectively, further improve the ability of deep learning systems to process and understand sign language in continuous contexts.

Hybrid models combine feature extraction, machine learning, and deep learning techniques to improve performance and robustness [30]. For instance, a hybrid model might integrate image-based CNNs with RNNs to capture a gesture sequence's spatial and temporal aspects. By combining multiple methods, these models can overcome the limitations of individual approaches, such as the inability of traditional machine learning algorithms to handle complex, dynamic gestures. Hybrid models are beneficial in real-world applications, where systems must process diverse input data, such as video and sensor data, to ensure accurate recognition across different environments and signers. Integrating multiple recognition methods helps create more adaptable, correct, and efficient systems that perform well in various practical scenarios.

In developing and accessing SLR systems, the choice of evaluation metrics plays a critical role in determining the effectiveness and efficiency of the models. These metrics quantify how well a system performs in terms of accuracy, reliability, speed, and robustness, offering insight into its practical utility for real-world applications [31]. Key evaluation metrics for SLR include accuracy, precision, recall, F1-score, real-time recognition accuracy, the confusion matrix, and time complexity/latency.

Accuracy is one of the most commonly used evaluation metrics, representing the overall performance of the SLR system. It is calculated as the ratio of correctly predicted gestures to the total number of predictions made. In other words, accuracy measures the proportion of correct classifications relative to the total number of

samples [32]. While it provides a general sense of model performance, accuracy alone may not always be sufficient, especially in imbalanced datasets where certain classes (gestures) are more frequent than others. In such cases, accuracy can be misleading, as a model might achieve high accuracy by simply predicting the majority class while neglecting the minority class.

In addition to accuracy, precision, recall, and F1-score are often used to provide a more detailed assessment of the model's performance, particularly in scenarios where class imbalance or varying costs of false positives and false negatives are a concern [33]. Precision measures the proportion of accurate optimistic predictions (correctly identified gestures) of all instances predicted as positive. In the context of SLR, high precision indicates that the system does not frequently misclassify non-sign gestures as valid signs. On the other hand, Recall quantifies the proportion of accurate optimistic predictions out of all actual positive instances, providing insight into how well the system identifies all possible correct gestures. Recall is particularly useful when the goal is to ensure that no valid sign is missed. The F1-score is the harmonic mean of precision and recall, offering a balanced measure of both metrics. It is particularly valuable in scenarios where precision and recall must be balanced, as it considers false positives and negatives. The F1 score provides a more nuanced evaluation than accuracy, especially in datasets with varying gesture frequencies.

Real-time recognition accuracy is another critical metric, especially for applications that require immediate feedback, such as human-computer interaction or real-time translation [34], [35]. It measures how well the SLR system performs in a live, interactive setting, where the system needs to process and classify gestures within strict time constraints. Real-time accuracy evaluates both the accuracy of gesture recognition and the time it takes to produce results [36]. It is an essential metric for assessing the system's feasibility in practical applications, as slow processing times may hinder the user experience, even if the model is otherwise highly accurate.

The confusion matrix is a comprehensive tool used to evaluate the performance of classification models in a more granular way. It provides a table that summarizes the number of true positives, false positives, true negatives, and false negatives for each class, offering insights into how well the model distinguishes between different gestures. The confusion matrix can help identify common misclassifications, revealing patterns such as sure signs being frequently confused with one another [37]. This diagnostic tool is handy for fine-tuning models, identifying areas of improvement, and understanding specific weaknesses in gesture recognition.

Time complexity and latency are important metrics when evaluating the efficiency of the recognition system. Time complexity refers to the computational cost associated with processing the input data, typically expressed in the number of operations required as a function of the input size. In the context of SLR, time complexity is a key consideration when the system needs to process large volumes of data, such as high-resolution video or 3D sensor inputs. A model with high time complexity may become impractical for real-time or large-scale applications. Latency, on the other hand, refers to the time delay between the input of a gesture and the system's response or output. In real-time SLR systems, low latency ensures users receive prompt feedback on their gestures [38]. High latency can lead to a suboptimal user experience, especially in interactive scenarios such as sign language translation or communication with assistive devices. Both time complexity and latency are crucial for evaluating the system's practical performance in real-world conditions, where rapid processing and responsiveness are often paramount.

One of the most impactful applications of SLR is in assistive technologies. For individuals who are deaf or hard of hearing, effective communication with non-signers is often a challenge, particularly in environments where interpreters are not readily available [39]. SLR systems can bridge this communication gap by translating sign language into spoken or written language in real time, allowing users to communicate more effectively in public spaces, workplaces, healthcare settings, and educational institutions [40]. For instance, real-time sign language translation tools, such as mobile apps or dedicated devices, enable deaf individuals to interact with hearing people, thus improving social inclusion and reducing isolation. Additionally, SLR systems can be integrated into devices like smartphones, smart glasses, or smart home systems, enhancing the accessibility of everyday technologies for deaf users. By incorporating SLR into assistive devices, individuals with hearing impairments can engage with technology, receive notifications, and even control their environment using sign language gestures.

Human-computer interaction (HCI) is another significant area where SLR has considerable potential. As technology becomes increasingly intuitive and multimodal, there is a growing demand for interaction methods beyond traditional input devices like keyboards and touchscreens. SLR offers a natural, gesture-based mode of communication with computers, enabling users to interact with digital systems using sign language. This is particularly beneficial for deaf and hard-of-hearing individuals, who may find voice-based or text-based

interactions less accessible or inefficient. For instance, sign language-based interfaces could allow users to navigate digital environments, interact with virtual assistants, or control smart devices through hand gestures. Incorporating SLR into HCI not only enhances accessibility for the deaf community but also enables more inclusive user experiences for people of all abilities. Furthermore, sign language recognition could be integrated into immersive technologies, such as virtual and augmented reality, allowing users to perform natural, gesture-based interactions within these environments [41].

SLR systems are also being increasingly adopted in sign language learning tools, which serve as educational resources for learners and platforms for preserving sign language in digital formats. Learning sign language can be challenging, particularly for individuals who do not have direct exposure to the language or community. SLR-based tools provide an interactive and dynamic way to facilitate learning by offering real-time feedback on learners' sign production [42]. For example, these tools can assess whether the learner is making a sign correctly based on hand shape, orientation, and movement, thereby providing immediate corrective feedback. This helps learners improve their skills more efficiently and engagingly than traditional methods. Additionally, such tools can be used for language preservation, particularly for minority or endangered sign languages. By creating large datasets of sign language gestures and making them accessible digitally, SLR systems contribute to the documentation and standardization of sign language, ensuring its survival for future generations. Moreover, SLR-based applications can be integrated into educational institutions, offering students and educators a convenient and accessible medium for learning and teaching sign language.

The paper's contribution lies in its comprehensive exploration and analysis of different approaches, techniques, and datasets used for alphabet SLR worldwide. The paper examines various sign languages' recognition systems and their effectiveness in recognizing alphabet-based signs, highlighting patterns and methodologies that have emerged across different countries and cultures. It provides an in-depth review of the state-of-the-art SLR systems, identifying existing approaches' strengths, challenges, and limitations. Additionally, the study offers insights into the evolution of SLR technologies. It presents recommendations for future study guidelines, aiming to advance these systems' accuracy, efficiency, and applicability on a global scale. Overall, the paper serves as an important resource for researchers in the SLR field, contributing to the advancement of inclusive communication systems for the community of deaf and hard-of-hearing individuals.

As technology continues to advance, the integration of pattern recognition methods for recognizing sign languages has become increasingly prominent. These methods significantly enhance accessibility and interaction for deaf and hard-of-hearing individuals by utilizing sophisticated algorithms and machine learning techniques. These methods can analyze visual inputs—such as hand gestures and facial expressions—enabling real-time interpretation of sign language [43]. This technological progress not only streamlines communication between signers and non-signers but also supports the development of applications and devices that promote inclusivity, such as automatic translation tools and interactive learning platforms. By making sign language more accessible, these innovations foster greater understanding and collaboration across diverse communities.

2. RELATED WORKS

Sign language recognition has been a focus of research for years, with many studies examining ways to close communication gaps between the hearing-impaired community and the hearing population. A significant challenge in this field is the creation of precise and effective systems that can interpret hand gestures in real time. Several methodologies have been employed to recognize sign language, with deep learning techniques emerging as the most prominent approach.

The data collection process for this study involved analyzing research articles and publicly available datasets to identify relevant studies on SLR. Research articles were selected based on their use of datasets from widely recognized sign languages, such as ASL, BSL, and ArSL, with a focus on studies employing CNN, RNN, or their combinations. Key details about datasets, such as sample size, signer diversity, and gesture types (static or dynamic), were extracted to inform data preparation. The collected data was standardized through preprocessing, including resizing, normalization, and data augmentation techniques such as rotation, scaling, and flipping. For video data, frame extraction and temporal alignment ensured consistency in representing dynamic gestures.

The curated data was processed using a hybrid CNN-RNN architecture. CNN layers extracted spatial features like hand shapes and orientations, and RNN layers captured temporal dependencies in dynamic gestures [44]. The model was trained with optimized hyperparameters and evaluated against benchmarks reported in the research articles using metrics such as accuracy, precision, recall, and F1-score. To test real-world applicability, latency, and computational efficiency were assessed on edge

devices like Raspberry Pi. Cross-validation ensured robust performance, demonstrating the system's ability to generalize across diverse scenarios while addressing challenges in dataset diversity and real-time applications highlighted in prior studies.

One of the early focuses in the field has been the recognition of sign language alphabets. Table 1 shows various methodologies and models used in sign language recognition, with differing performance outcomes. For instance, the ASL alphabet has been effectively recognized using CNNs, a technique for efficiently classifying image data. Previous studies have highlighted the strong performance of CNNs in recognizing the ASL alphabet. This success is primarily attributed to CNNs' capability to capture intricate features from images, making them particularly effective for sign language recognition tasks. These advancements have spurred ongoing research on enhancing the models' performance and optimizing them for real-time applications.

Table 1. Comparison of Approaches for Sign Language Recognition: Methodologies, Models, and Performance

Study	Methodology	Dataset	Model	Accuracy	Challenges	Key Contributions
Hassanin, (2023). [45]	Fast Gradient Sign Method (FGSM) with Keras and TensorFlow	Arabic manuscript dataset	CNN-based Region Proposal Algorithm for object detection	99%	Handling multilingual and varied document categories	Proposed a new framework integrating adversarial training and RoI detection
Batool, (2022). [46]	EfficientNet models for lightweight deep learning	Arabic Sign Language (ArSL) gestures	Lightweight deep learning model using EfficientNet-Lite	94% classification accuracy	Designing lightweight models suitable for mobile devices	Achieved high classification accuracy with reduced computational requirements
Benjamin, (2023). [47]	Comparative analysis of machine learning models	Custom dataset of 24,300 images of Norwegian Sign Language alphabet signs	Support Vector Machine (SVM), CNN, and K-Nearest Neighbor (KNN)	SVM and CNN achieved 99.9% accuracy	Need for efficient and accurate models for NSL	Demonstrated effective models (SVM and CNN) for NSL recognition with high accuracy
Shagun, (2022). [26]	Bag of Visual Words (BOVW), SURF, SVM, CNN	Indian Sign Language alphabets include digits (0–9) and (A-Z)	SVM and CNN for classification	No specific accuracy was mentioned Predictions as text and speech	Effective segmentation of hand signs amidst varying backgrounds	Provides a real-time recognition system with text and speech output.
Itsaso, (2021). [48]	Hand landmarks extraction, Common Spatial Patterns (CSP), feature extraction	LSA64 dataset (Argentinian Sign Language)	Random Forest (RF), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP)	Accuracy between 0.90 and 0.95 on 42 signs	Communication barrier for the deaf and hard-of-hearing individuals without interpreters	Argentinian Sign Language recognition scheme using hand landmarks and CSP for feature extraction
Diponkor, (2021). [11]	Convolutional Neural Network (CNN), dataset pre-processing (normalization)	Sign Language MNIST (34,627 images, 27,455 training, 7,172 testing)	CNN for ASL (American Sign Language) alphabet recognition	99.78% accuracy on unseen data	Considerable inter-class variation in sign language, complexity of recognition	Using CNN to achieve high accuracy in recognizing ASL alphabets
Tzeico, (2024). [49]	Convolutional Neural Network (CNN), MediaPipe framework	336 test images of Mexican Sign Language (MSL) dactylological alphabet	CNN for static sign language recognition	84.57% accuracy, 83.33% sensitivity, 99.17% specificity (for letter samples)	Recognizing static signs with a limited dataset	Real-time classification of MSL on low-consumption equipment, improving accessibility
Jungpil, (2021). [24]	MediaPipe hands algorithm, feature extraction	American Sign Language Alphabet dataset	SLR based on vision, utilizing hand images captured by a webcam	99.39% (Massey), 87.60% (ASL Alphabet), 98.45% (Finger Spelling A dataset)	Achieving high accuracy with a vision-based approach and web camera	Cost-effective, inexpensive ASL recognition without the need for sensors, outperforming previous studies
Kinanti, (2024). [50]	Convolutional Neural Network (CNN) for finger spelling gesture recognition	BISINDO finger spelling gestures	CNN-based gesture recognition system, real-time prediction	97.5% accuracy	Inaccurate recognition of one letter, limited dataset size	Development of a BISINDO finger spelling gesture recognition system with high accuracy and real-time prediction
Zaraan, (2022). [51]	Deep learning-based real-time Arabic Sign Language Alphabet (ArSLA) recognition	scientific ArSLA dataset	AlexNet architecture	94.81% accuracy in real-time recognition	Not specified in detail, but real-time recognition in general	Development of a real-time ArSLA recognition model using AlexNet, achieving high accuracy

In the context of ArSL, a similar approach was taken, utilizing deep learning architectures like

AlexNet to achieve high accuracy in recognizing the ArSLA (Arabic Sign Language Alphabet) [51]. This study showed that the best-performing deep learning architecture for real-time recognition was AlexNet, with a reported accuracy of 94.81%. The research highlighted the importance of selecting appropriate architectures for real-time applications, particularly for ArSL, where sign language is represented by specific signs or fingerspelling. This effort to create a real-time system for ArSLA recognition contributes significantly to the field, as it addresses the unique challenges posed by the Arabic script and sign language structure [52].

Another significant contribution comes from studies focusing on Indian Sign Language (ISL), which has not received as much attention as ASL. Researchers developed a Bag of Visual Words (BOVW) model to classify ISL alphabets and digits in a live video stream [26]. This approach utilized a Support Vector Machine (SVM) and CNN for classification, achieving high accuracy for alphabets and digits, with CNNs outperforming other classifiers. Notably, this work contributed to the recognition of ISL in real-time, with a system that outputs both text and speech, providing an interactive and accessible solution for the hearing-impaired community.

For Mexican Sign Language (MSL), using CNNs was also explored to recognize static signs from video frames. The study employed the MediaPipe framework to detect hand landmarks and used these landmarks as input features for a CNN-based model. The results showed an accuracy of 83.63%, highlighting the potential of CNNs for real-time MSL recognition [49]. The system also demonstrated high specificity, with the ability to classify signs even in varied conditions, such as different backgrounds, suggesting that CNNs can handle the diversity of hand shapes and positions inherent in sign language gestures.

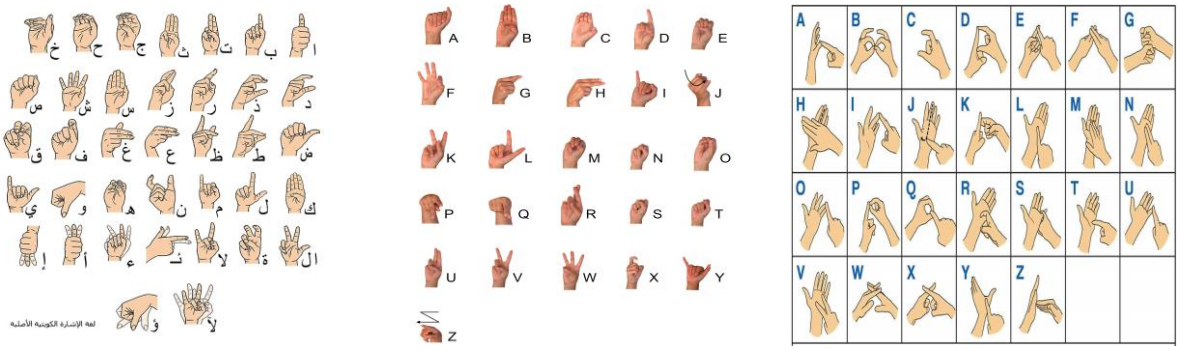
Finally, a recent study on Argentinian Sign Language (LSA) recognition leveraged hand landmarks and the Common Spatial Patterns (CSP) algorithm to improve the classification of signs from the LSA64 dataset. This method incorporated various classifiers like K-Nearest Neighbors (KNN), Random Forest (RF), and Multilayer Perceptron (MLP) to achieve accuracy rates varying between 0.90 to 0.95 [48]. The study's use of hand landmark extraction, combined with CSP for dimensionality reduction, represents an innovative way to address challenges in sign language recognition by improving feature extraction and classifier performance.

3. METHODOLOGIES IN SIGN LANGUAGE RECOGNITION

In SLR, CNNs have become one of the most popular methodologies, especially for static gesture recognition. CNNs are effective because they autonomously learn and extract features from raw image data, making them compatible with analyzing hand shapes, positions, and orientations. This method is beneficial when the sign language gestures are represented as images, as CNNs can accurately classify these gestures. By processing the data through multiple layers, CNNs capture hierarchical patterns crucial for distinguishing hand shapes and movements in sign language.

RNNs are commonly used for dynamic gestures, which involve movements over time. They are highly effective at processing sequential data by recognizing temporal relationships, making them ideal for recognizing gestures that change over time. Specifically, Long-Short-Term Memory (LSTM) networks, a kind of RNN, are often employed in sign language recognition [53]. LSTMs can remember previous states over long sequences, making them more effective at interpreting continuous gestures and complete sentences [54]. Combining CNNs for feature extraction besides RNNs for sequence prediction allows for improved recognition of static and dynamic signs, enhancing the general performance of SLR systems.

CNNs and RNNs are two robust models commonly used in SLR, each specializing in handling different aspects of sign language datasets. CNNs are particularly adept at processing spatial features, making them ideal for recognizing static gestures like hand shapes and orientations. When applied to datasets such as ASL, BSL, and ArSL, as shown in Figure 2, CNNs can effectively learn the visual characteristics of each sign. These datasets typically include images representing individual signs or letters of the alphabet. For instance, each letter in ASL or BSL, or the individual gestures in ArSL, can be captured in a single image frame, which CNNs process through convolutional layers to identify key features such as shapes, edges, and textures. This enables CNNs to classify each gesture based on its unique spatial attributes.



(ArSL-A) (ASL-A) (BSL-A)

Figure 2. Visual Representation of Arabic (ArSL-A), American (ASL-A), and British (BSL-A) Sign Language Alphabets

Conversely, RNNs, particularly Long Short-Term Memory (LSTM) networks, are considered to handle sequential data and are well-suited for dynamic sign language recognition, where gestures involve motion and change over time [29]. Sign language often consists of sequences of gestures or continuous signs, which RNNs can understand due to their ability to learn temporal dependencies between frames in a video. In datasets like ASL, BSL, and ArSL, RNNs track the evolution of hand gestures from one frame to the next, understanding the flow and progression of movements in a sign language sentence or phrase. For example, the transition from one letter to another in a spelling sequence or the motion of a hand for a common phrase would be captured and processed by an RNN, enabling it to recognize the sequence of gestures.

Example of SLA Datasets:

1) ASL Dataset: The American Sign Language dataset contains hand gestures for each letter in the ASL alphabet and often includes static images for individual letters and dynamic sequences for entire phrases or sentences. This dataset trains models to recognize hand gestures as part of English, where a specific gesture represents each letter or word.
2) BSL Dataset: Similar to the ASL dataset, the British Sign Language dataset includes gestures for the letters of the BSL alphabet. It may also contain gestures for common words and phrases in the UK. Like ASL, BSL signs are captured in both static images and video sequences to help train models to recognize letters and full expressions [55].
3) ArSL Dataset: The Arabic Sign Language dataset includes hand gestures representing the Arabic alphabet and some common phrases [56]. ArSL is used across Arabic-speaking regions so that the dataset may contain variations in the signs based on regional dialects. Like the other two datasets, it includes static and dynamic sign representations.

CNNs are adequate for visual recognition tasks, such as identifying hand gestures in sign language [57]. CNNs automatically learn hierarchical features from images, making them ideal for recognizing patterns in sign language gestures. The process can be broken down into key steps, from preprocessing the input images to making estimates based on learned features. Figure 3 is a flowchart that illustrates the general process of how CNNs are applied in SLR.

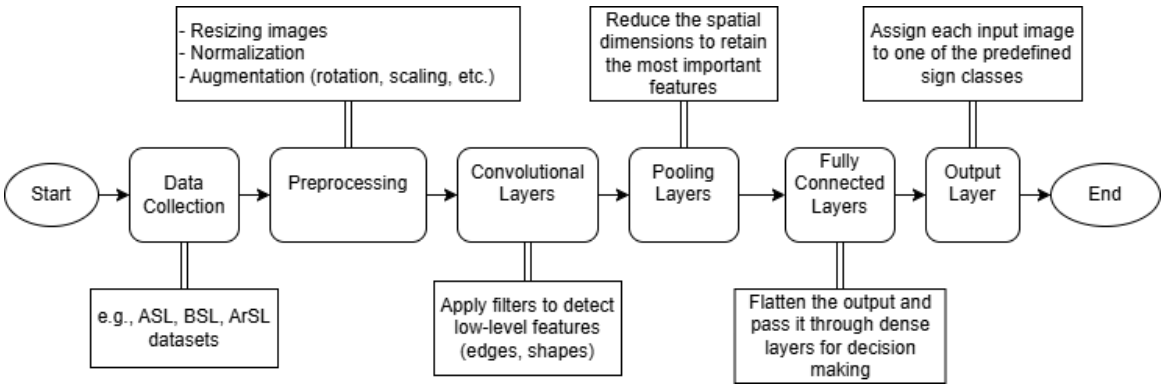


Figure 3. The Flowchart of CNN to recognize sign language

The first step in applying CNNs to SLR is the collection of a suitable dataset. This dataset typically includes images or video frames representing different sign language gestures, such as those from

ASL, BSL, or ArSL [58]. Each image or frame in the dataset corresponds to a specific sign or alphabet, and the goal is to train the CNN to recognize these signs. The diversity in the dataset, such as variations in hand orientations, shapes, lighting conditions, and backgrounds, plays a crucial role in training a model that can generalize well to real-world scenarios [59].

After collecting the data, the images must be preprocessed to make them suitable for input into the CNN. Preprocessing typically involves resizing the images to a uniform size to ensure consistency across the dataset. Since CNNs work more efficiently with normalized data, the pixel values of each image are usually scaled to a range between 0 and 1. This normalization helps the network converge faster during training. Furthermore, data augmentation procedures, such as rotating, flipping, and scaling images, are applied to artificially increase the dataset's size and variety. These techniques help the model generalize better and avoid overfitting, especially when training on smaller datasets.

The core of a CNN lies in its convolutional layers, which are responsible for extracting features from the input images. In these layers, filters (or kernels) are applied to the image, detecting low-level features such as edges, textures, and corners [60]. Each filter produces a feature map, highlighting different aspects of the image. For example, one filter may capture the outlines of a hand, while another could detect the shape of a finger. As the image moves through successive convolutional layers, the network develops progressively more abstract representations, allowing it to recognize more complex patterns critical for SLR, such as the shape and position of the hands.

After the convolutional layers, pooling layers are used to reduce the spatial dimensions of the feature maps. Pooling serves to down-sample the information while retaining the most important features. This step helps reduce computational cost and decreases the risk of overfitting. Typically, max pooling is used, where the highest value from a group of neighboring pixels is retained. Pooling allows the network to focus on the most salient features of the image, such as the key points of hand gestures, and discard less important details. This also helps the network become more robust to variations like translation and minor image distortions.

Once the feature maps have been pooled and reduced in size, the data is flattened into a one-dimensional vector and passed through fully connected layers. These layers are responsible for interpreting the features and making the final classification decision. The fully connected layers map the extracted features to specific classes, such as different sign language symbols or letters. Each neuron in the fully connected layer corresponds to a potential output class, and the network learns to associate the feature vector with the correct sign language gesture [61]. This step is crucial as it combines the high-level features extracted by the convolutional layers and prepares them for classification.

The output layer of the CNN provides the final prediction for the input image [62]. Based on the learned features and the classification performed by the fully connected layers, the output layer assigns a probability to each potential class. Typically, a softmax activation function is used, which outputs a probability distribution over all possible classes [63]. The class with the highest probability is selected as the model's prediction. In the case of SLR, this could be the letter or word represented by the input gesture. For example, the model may output "A" if it identifies the gesture as the ASL sign for the letter "A."

RNNs are a class of neural networks that handle sequential data, making them compatible with speech recognition, language modeling, and sign language recognition from video frames [64]. Unlike CNNs, which specialize in spatial patterns in images, RNNs can learn temporal dependencies, making them ideal for interpreting sequences of frames or videos in sign language recognition [65].

Figure 4 shows a flowchart of the general process when using RNN for sign language recognition, illustrating the sequence from input frames to final prediction. The process includes steps such as frame extraction, feature extraction, sequence input into the RNN, and output prediction of the sign language gesture.

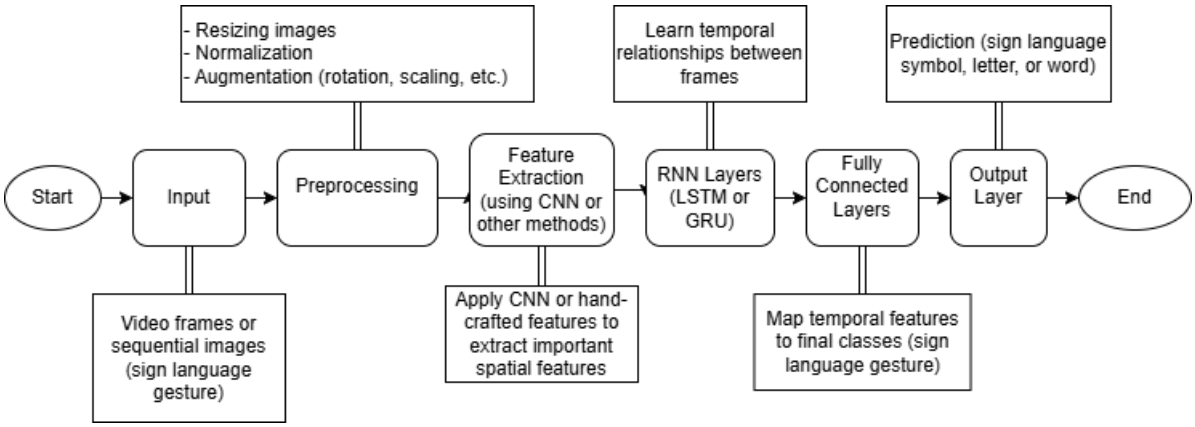


Figure 4. The Flowchart of RNN to recognize sign language

1) Input Sequence: The initial step in the RNN-based SLR model is to capture the sequence of frames. These frames are typically extracted from a video where the sign language gesture is performed. Each frame corresponds to a hand gesture at a specific time. The input could be a series of images or video frames representing the movement of the hands. These frames are collected and arranged in the correct temporal order.
2) Preprocessing: Before feeding the frames into the RNN, preprocessing steps are performed to prepare the data for the model. This phase includes resizing the images, normalizing the pixel values, and augmenting the dataset by applying rotation, zooming, or shifting transformations to increase the variety of inputs and improve the model's robustness. Normalization ensures that all pixel values lie within the same range, which helps the model to learn more effectively [66].
3) Feature Extraction: RNNs alone cannot effectively process raw pixel data. To address this limitation, feature extraction is performed before feeding the data into the RNN. CNNs are commonly used in this step to extract spatial features from each frame. CNNs help identify hand shapes, positions, orientations, and other spatial patterns critical for understanding gestures. Alternatively, hand-crafted features such as hand landmarks or motion descriptors may also be used.
4) Sequence Input to RNN: After the spatial features are extracted from each frame, the sequence of frames is fed into the RNN [67]. Each frame's features are passed into the RNN, and the network processes them in order, considering both the current and past frames in the sequence. This is where RNNs excel, as they have a memory mechanism that permits them to maintain information from former frames, allowing them to learn the temporal relationships between frames.
5) RNN Layers (LSTM or GRU): The core of the RNN is composed of specialized layers, for instance, Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRU). These units are designed to capture long-term dependencies in the sequential data, which is crucial for SLR. LSTM and GRU networks help mitigate the vanishing gradient problem and allow the model to retain important information over many time steps. For example, an RNN can learn how a hand shape changes over time and interpret how these changes contribute to the meaning of the sign.
6) Fully Connected Layer: Once the RNN has processed the entire sequence of frames and learned the temporal patterns, its output is passed through one or more fully connected layers. These layers aggregate the features learned by the RNN and map them to the final output classes, which correspond to different signs, letters, or symbols in the sign language alphabet. The fully connected layer takes the learned temporal features from the RNN and produces the final prediction.
7) Output: The output layer typically generates the final prediction using a softmax activation function. The softmax function turns the output into a probability distribution for the potential classes. The class with the highest probability is selected as the predicted sign language gesture. Depending on the complexity of the model, the output could represent a specific letter, word, or complete gesture.

CNNs and RNNs are extensively used in SLR but differ significantly in their approach to processing data and handling the complexity of sign language gestures. CNNs are primarily designed to process spatial information in images. They apply convolutional layers to extract spatial features from images of hand gestures, allowing the model to learn local patterns like hand shapes, orientation, and

positioning. This makes CNNs particularly effective for recognizing individual signs based on visual features from static images or video frames[68].

On the other hand, RNNs are considered to handle sequential data, making them more appropriate for recognizing dynamic gestures in sign language. Unlike CNNs, which process individual frames or images independently, RNNs consider the temporal sequence of frames in a video. This enables RNNs to model the motion and changes over time, capturing the transitions between positions, hand movements, and gestures. RNNs are thus ideal for understanding the context of a sign language gesture, which often involves a combination of hand shape, movement, and location, with crucial temporal dependencies for accurate recognition.

While CNNs excel at extracting spatial features from individual frames, they do not inherently capture the temporal relationships between frames. In contrast, RNNs, with their built-in memory mechanism (like LSTM or GRU), can retain information from previous frames, making them more adept at processing continuous sequences, such as a series of hand movements. This makes RNNs more suitable for recognizing dynamic and context-dependent gestures in sign language, where the meaning is often derived from the entire sequence of actions rather than a single frame.

In many sign language recognition systems, CNNs and RNNs are combined to take advantage of the benefits of both models. CNNs are initially employed to extract spatial features from each frame, and then RNNs are used to process these features over time, capturing the dynamics of the gesture. This combination allows for a more robust recognition system that can handle both the visual complexity of static hand shapes and the temporal complexity of dynamic movements [68].

4. DISCUSSION

Sign language involves dynamic gestures, trajectory properties, and multidimensional feature vectors, making recognizing it challenging. Despite these complexities, researchers are focused on developing generalized, reliable, and robust SLR models. Incorporating multidimensional features is an emerging approach that has shown promise in enhancing recognition accuracy.

This review paper seeks to provide a clear understanding and practical guidance to the SLR research community. Developing effective SLR models to support the hand-signing community remains a prominent research area within pattern recognition, computer vision, and natural language processing.

The limitations of current datasets and their sizes present significant challenges for SLR. One major issue is the ambiguity and lack of comprehensive training datasets, which makes SLR systems vulnerable to errors. Large-scale and standardized datasets that include manual and non-manual features are essential for effective SLR. However, the current datasets often face barriers due to inadequate recording, collection, and measuring equipment, leading to reduced performance.

Several factors impact the quality of the datasets, including poor camera quality, improper camera setups, and issues with multi-camera synchronization. When camera resolution is low, the clarity of signs is compromised, decreasing the system's accuracy. Similarly, improper camera setup can result in losing important sign information, especially when signs are dynamic or static. If multiple cameras are used, a lack of synchronization may cause information loss, further degrading performance. Additionally, the devices' reliability, cost-effectiveness, and ease of maintenance are critical for consistent performance. The environment in which data is captured also plays a crucial role. Background noise, improper lighting, and poor illumination can all negatively impact the dataset's quality, leading to misclassifications and reduced recognition rates. The distance between the camera and the signer must also be optimal; too close or too far can significantly affect the system’s performance.

The current trends in SLR face several limitations, which hinder their accuracy and performance. One major issue is the variability between different signers, affecting recognition. For instance, variations in the speed and continuity of signs make segmentation and feature extraction more challenging. Additionally, occlusions, such as hand-to-face or hand-to-hand overlap, and factors like long sleeves or colored gloves can obstruct sign recognition. There is also considerable variation in how signs are performed by different individuals, which complicates the process.

Video-related challenges are another limitation, as handling video data often exceeds the capacity of limited GPU memory. Since many CNN techniques are image-based, videos, with their additional temporal dimension, pose problems. A simple resizing process can result in the loss of important temporal information, affecting the fine-tuning and classification of each frame. Network design challenges also impact performance, with location and illumination influencing recognition and classification abilities. Moreover, choosing the right batch size during training is critical, as a larger batch size can reduce local convergence, while smaller batch sizes increase training costs. The selection of appropriate loss functions and optimal hyperparameters also presents hurdles. Despite the advancements in deep learning networks, which

have enhanced SLR accuracy, these limitations remain significant, and addressing them is crucial for further development in SLR.

SLR has many potential applications, particularly when integrated with human-computer interaction. One key application is virtual reality (VR), where users can experience artificial simulations of the real world, using SLR to communicate with virtual environments. In smart homes, SLR can be used to monitor, access, and control household devices through sign language, enhancing the accessibility of innovative technologies. In healthcare, SLR can assist patients, enabling better communication between patients and healthcare providers and improving the quality of life and healthcare services. Furthermore, SLR can be crucial in social safety, ensuring safe interactions, and minimizing social threats for individuals with hearing impairments. In telehealth, it allows for remote consultations, making healthcare services more accessible.

SLR also has applications in virtual shopping, providing a more inclusive and comfortable shopping experience by enabling customers to use sign language to interact with virtual stores. In digital signatures, SLRs can authenticate information through electronic signs. The gaming and entertainment industry can benefit from SLR by providing users with a more immersive and interactive experience through sign language-based controls. In text and voice assistance, SLR can be combined with speech recognition to offer better communication, allowing users to interact through text and sign language. Education is another domain where SLR can be essential, facilitating enhanced learning and communication for students with hearing impairments. Moreover, one notable and impactful application of SLR could be in reciting verses from the Al-Quran [69]. By enabling individuals to recite Quranic verses in sign language, SLR can promote inclusivity and enhance the learning experience for the deaf and hard-of-hearing community, offering a meaningful way to engage with the sacred text.

5. RESEARCH SCOPE AND FUTURE DIRECTION

Compared to recent advancements in automatic speech recognition, SLR is still in its primary stages of development and lags significantly behind. Much research has been conducted in the field of SLR, with numerous studies focused on achieving higher performance using advanced techniques such as deep learning, machine learning, optimization, and experimentation with advanced hardware and sensors. Despite these efforts, several challenges remain unresolved. These include issues with distinctiveness and handling sign variations, difficulties related to the fusion of multiple sensors or cameras, managing multi-modal data, computational challenges, maintaining consistency, and effectively handling large vocabularies. Additionally, there is a need for standard datasets that can be used universally for training and evaluation.

Future directions for SLR include the need for a better understanding of optimal hyperparameter estimation strategies for SLR model design. Most current models are developed using controlled lab-based datasets, so building models that can function effectively in uncontrolled environments is a crucial area of focus. Another significant challenge lies in designing user-friendly, realistic, and robust sign language models that can be widely adopted. The development of high-precision capturing devices such as sensors and cameras and novel training strategies to reduce computational complexities will also play a key role in the progress of the field. Lightweight CNNs for SLR are an area of active research, along with integrating multi-modal data to enhance recognition accuracy [70]. Researchers are also striving to create a generic, automatic SLR model that can be applied across different sign languages and contexts. This review paper serves as a comprehensive guide for researchers, outlining the challenges, gaps, and future research directions in SLR, with the ultimate goal of developing innovative models that can assist the hand-talking community and contribute to social well-being.

6. CONCLUSION

In conclusion, this research paper highlights the significant progress made in the SLR field, focusing on applying advanced machine learning techniques such as CNN and RNN to improve accuracy and efficiency in recognizing sign languages. The integration of these deep learning models has shown promising results in overcoming the challenges associated with the dynamic, multi-dimensional, and often ambiguous nature of sign language gestures. However, despite the advancements, several obstacles remain, such as the limitations of current datasets, the need for standardized and large-scale datasets, and the complexity involved in handling various sign languages with distinct characteristics.

Furthermore, this paper has discussed the potential applications of SLR, particularly its integration with human-computer interaction in diverse fields like healthcare, education, and social safety. It also emphasized the importance of developing realistic, user-friendly systems that work in uncontrolled environments, which remains a significant challenge for future SLR models. The

research has underscored the importance of addressing the computational and data limitations, such as handling large vocabularies, achieving high recognition accuracy, and ensuring consistency across different signers and environments. Lastly, the future of SLR looks promising, with the need for continued innovation in model design, dataset development, and the integration of multi-modal systems to create more robust, real-time solutions that can benefit the hand-talking community and enhance accessibility across various sectors of society.

ACKNOWLEDGMENT

I want to express my deepest gratitude to Prof. Muchlas, Prof. Imam Riadi, Prof. Abdul Fadhil, and Prof. Tole Sutikno for their invaluable support and insightful advice throughout the writing of this research article. Their guidance and encouragement have been instrumental in shaping this work, and I am sincerely thankful for their dedication and expertise.

BIBLIOGRAPHY

[1] M. D. Meitantya, C. A. Sari, E. H. Rachmawanto, and R. R. Ali, "VGG-16 Architecture on CNN for American Sign Language Classification," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 1165–1171, Jul. 2024, doi: 10.52436/1.jutif.2024.5.4.2160.

[2] A. Yudhana, J. Rahmawan, and S. A. Akbar, "EonTex Conductive Strechtable Sensor Response on Smart Glove for Sign Language," 2019.

[3] G. Tharwat, A. M. Ahmed, and B. Bouallegue, "Arabic Sign Language Recognition System for Alphabets Using Machine Learning Techniques," *Journal of Electrical and Computer Engineering*, vol. 2021, 2021, doi: 10.1155/2021/2995851.

[4] A. Duarte *et al.*, "How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language," 2021. [Online]. Available: http://how2sign.github.io/

[5] C. Kwenda, M. Gwetu, and J. V. Fonou-Dombeu, "Ontology with Deep Learning for Forest Image Classification," *Applied Sciences (Switzerland)*, vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13085060.

[6] S. Luong, "Video Sign Language Recognition using Pose Extraction and Deep Learning Models," San Jose State University, San Jose, CA, USA, 2023. doi: 10.31979/etd.jm4c-myd4.

[7] Z. R. Saeed, Z. B. Zainol, B. B. Zaidan, and A. H. Alamoodi, "A Systematic Review on Systems-Based Sensory Gloves for Sign Language Pattern Recognition: An Update from 2017 to 2022," 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3219430.

[8] G. Latif, N. Mohammad, J. Alghazo, R. AlKhalaf, and R. AlKhalaf, "ArASL: Arabic Alphabets Sign Language Dataset," *Data Brief*, vol. 23, Apr. 2019, doi: 10.1016/j.dib.2019.103777.

[9] A. A. J. Jim, I. Rafi, M. Z. Akon, U. Biswas, and A. Al Nahid, "KU-BdSL: An open dataset for Bengali sign language recognition," *Data Brief*, vol. 51, Dec. 2023, doi: 10.1016/j.dib.2023.109797.

[10] I. D. A. Rachmawati, R. Yunanda, M. F. Hidayat, and P. Wicaksono, "Deep Transfer Learning for Sign Language Image Classification: A Bisindo Dataset Study," *Engineering, MAthematics and Computer Science Journal (EMACS)*, vol. 5, no. 3, pp. 175–180, Sep. 2023, doi: 10.21512/emacsjournal.v5i3.10621.

[11] M. A. Hossain, D. Bala, B. Sarkar, and I. Abdullah, "American Sign Language Alphabets Recognition using Convolutional Neural Network." [Online]. Available: https://www.researchgate.net/publication/352878275

[12] B. Alsharif, A. S. Altaher, A. Altaher, M. Ilyas, and E. Alalwany, "Deep Learning Technology to Recognize American Sign Language Alphabet," *Sensors*, vol. 23, no. 18, Sep. 2023, doi: 10.3390/s23187970.

[13] I. Irvanizam, I. Horatius, and H. Sofyan, "Applying Artificial Neural Network Based on Backpropagation Method for Indonesian Sign Language Recognition," *International Journal of Computing and Digital Systems*, vol. 14, no. 1, pp. 975–985, 2023, doi: 10.12785/ijcds/140176.

[14] Z. Zhang *et al.*, "Artificial intelligence in cyber security: research advances, challenges, and opportunities," *Artif Intell Rev*, vol. 55, no. 2, pp. 1029–1053, Feb. 2022, doi: 10.1007/s10462-021-09976-0.

[15] A. W. Salehi *et al.*, "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," Apr. 01, 2023, *MDPI*. doi: 10.3390/su15075930.

[16] B. Joksimoski *et al.*, "Technological Solutions for Sign Language Recognition: A Scoping Review of Research Trends, Challenges, and Opportunities," 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3161440.

[17] M. S. Amin, S. T. H. Rizvi, A. Mazzei, and L. Anselma, "Assistive Data Glove for Isolated Static Postures Recognition in American Sign Language Using Neural

Network," *Electronics (Switzerland)*, vol. 12, no. 8, Apr. 2023, doi: 10.3390/electronics12081904.

[18] A. Hekmat, M. Ali, H. H. Abbas, and I. Shahadi, "Sign Language Recognition and Hand Gestures Review," vol. 02, no. 04, 2022, [Online]. Available: https://kjes.uokerbala.edu.iq/

[19] A. Rahagiyanto, A. Basuki, R. Sigit, A. Anwar, and M. Zikky, "Hand Gesture Classification for Sign Language Using Artificial Neural Network," in *ICSEC 2017 - 21st International Computer Science and Engineering Conference 2017, Proceeding*, Institute of Electrical and Electronics Engineers Inc., Aug. 2018, pp. 205–209. doi: 10.1109/ICSEC.2017.8443898.

[20] M. Hocine, B. Mohammed Kamel, and B. Mohamed, "Hand gesture and sign language recognition based on deep learning."

[21] C. Lu, M. Kozakai, and L. Jing, "Sign Language Recognition with Multimodal Sensors and Deep Learning Methods," *Electronics (Switzerland)*, vol. 12, no. 23, Dec. 2023, doi: 10.3390/electronics12234827.

[22] A. Muis, S. Sunardi, and A. Yudhana, "CNN-based Approach for Enhancing Brain Tumor Image Classification Accuracy," *International Journal of Engineering, Transactions B: Applications*, vol. 37, no. 5, pp. 984–996, May 2024, doi: 10.5829/ije.2024.37.05b.15.

[23] B. Awaji *et al.*, "Hybrid Techniques of Facial Feature Image Analysis for Early Detection of Autism Spectrum Disorder Based on Combined CNN Features," *Diagnostics*, vol. 13, no. 18, Sep. 2023, doi: 10.3390/diagnostics13182948.

[24] J. Shin, A. Matsuoka, M. A. M. Hasan, and A. Y. Srizon, "American sign language alphabet recognition by extracting feature from hand pose estimation," *Sensors*, vol. 21, no. 17, Sep. 2021, doi: 10.3390/s21175856.

[25] M. S. Abdallah, E. Hemayed, M. S. Abdalla, and E. E. Hemayed, "Dynamic Hand Gesture Recognition of Arabic Sign Language using Hand Motion Trajectory Features Dynamic Hand Gesture Recognition of Arabic Sign Language using Hand Motion Trajectory Features," 2013. [Online]. Available: https://www.researchgate.net/publication/258172682

[26] S. Katoch, V. Singh, and U. S. Tiwary, "Indian Sign Language recognition system using SURF with SVM and CNN," *Array*, vol. 14, Jul. 2022, doi: 10.1016/j.array.2022.100141.

[27] A. Z. Nugraha, R. F. Salsabila, A. N. Handayani, A. P. Wibawa, E. Hitipeuw, and K. Arai, "Decision tree based algorithms for Indonesian Language Sign System (SIBI) recognition," *Applied Engineering and Technology*, vol. 3, no. 2, pp. 86–101, Aug. 2024, doi: 10.31763/aet.v3i2.1536.

[28] S. Oulad-Naoui, H. Ben-Abderrahmane, A. Chagha, and A. Cherif, "An LSTM-based System for Dynamic Arabic Sign Language Recognition," in *Proceedings ofthe International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024)*, 2024, pp. 313–323. doi: 10.2991/978-94-6463-496-9_24.

[29] B. Sundar and T. Bagyammal, "American Sign Language Recognition for Alphabets Using MediaPipe and LSTM," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 642–651. doi: 10.1016/j.procs.2022.12.066.

[30] A. M. Buttar, U. Ahmad, A. H. Gumaei, A. Assiri, M. A. Akbar, and B. F. Alkhamees, "Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs," *Mathematics*, vol. 11, no. 17, Sep. 2023, doi: 10.3390/math11173729.

[31] N. Al Mudawi *et al.*, "Innovative healthcare solutions: robust hand gesture recognition of daily life routines using 1D CNN," *Front Bioeng Biotechnol*, vol. 12, 2024, doi: 10.3389/fbioe.2024.1401803.

[32] N. Adaloglou *et al.*, "A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition," *IEEE Trans Multimedia*, vol. 24, pp. 1750–1762, 2022, doi: 10.1109/TMM.2021.3070438.

[33] A. M. Buttar, U. Ahmad, A. H. Gumaei, A. Assiri, M. A. Akbar, and B. F. Alkhamees, "Deep Learning in Sign Language Recognition: A Hybrid Approach for the Recognition of Static and Dynamic Signs," *Mathematics*, vol. 11, no. 17, 2023, doi: 10.3390/math11173729.

[34] P. T. Nguyen, T. H. Nguyen, N. X. N. Hoang, H. T. B. Phan, H. S. H. Vu, and H. N. Huynh, "Exploring MediaPipe optimization strategies for real-time sign language recognition," *CTU Journal of Innovation and Sustainable Development*, vol. 15, no. ISDS, pp. 142–152, Oct. 2023, doi: 10.22144/ctujoisd.2023.045.

[35] A. Tayade and A. Halder, "Real-time Vernacular Sign Language Recognition using MediaPipe and Machine Learning," *International Journal of Research Publication and Reviews*, vol. 2, no. 5, 2021, doi: 10.13140/RG.2.2.32364.03203.

[36] S. K. Hussin, O. Mohamed, ; Mustafa Mohamed, E. Ahmed, and O. Mahmoud, "Real-time Arabic sign language translator Using media pipe and LSTM." [Online]. Available: https://plomscience.com/journals/index.php/PLOMSAI/index

[37] M. Papatsimouli, P. Sarigiannidis, and G. F. Fragulis, "A Survey of Advancements in Real-Time Sign Language Translators: Integration with IoT Technology," Aug. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/technologies11040083.

[38] M. S. Abdallah, G. H. Samaan, A. R. Wadie, F. Makhmudov, and Y. I. Cho, "Light-Weight Deep Learning Techniques with Advanced Processing for Real-Time Hand Gesture Recognition," *Sensors*, vol. 23, no. 1, Jan. 2023, doi: 10.3390/s23010002.

[39] R. S. Abdul Ameer, M. A. Ahmed, Z. T. Al-Qaysi, M. M. Salih, and M. L. Shuwandy, "Empowering Communication: A Deep Learning Framework for Arabic Sign Language Recognition with an Attention Mechanism," *Computers*, vol. 13, no. 6, Jun. 2024, doi: 10.3390/computers13060153.

[40] R. S. Abdul Ameer, M. A. Ahmed, Z. T. Al-Qaysi, M. M. Salih, and M. L. Shuwandy, "Empowering Communication: A Deep Learning Framework for Arabic Sign Language Recognition with an Attention Mechanism," *Computers*, vol. 13, no. 6, Jun. 2024, doi: 10.3390/computers13060153.

[41] A. Rakhmadi and A. Yudhana, "Virtual Reality and Augmented Reality in Sign Language Recognition: A Review of Current Approaches," *International Journal of Informatics and Computation (IJICOM)*, vol. 6, no. 2, 2024, doi: 10.35842/ijicom.

[42] T. H. Noor *et al.*, "Real-Time Arabic Sign Language Recognition Using a Hybrid Deep Learning Model," *Sensors*, vol. 24, no. 11, Jun. 2024, doi: 10.3390/s24113683.

[43] I. Puspitasari, A. Yudhana, D. E. Wati, and S. Al Irfan, "Recognizing Micro Expression Pattern Using Convolutional Neural Networks (CNN) Method During Emotion Regulation Training for Parents in The Pandemic Era," 2023. [Online]. Available: https://icistech.org/index.php/icistech/

[44] N. Nurhadi, E. A. Winanto, R. M. Said, J. Jasmir, and L. Afuan, "PATTERN CLASSIFICATION SIGN LANGUAGE USING FEATURES DESCRIPTORS AND MACHINE LEARNING," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 2, pp. 349–356, Apr. 2024, doi: 10.52436/1.jutif.2024.5.2.1228.

[45] H. M. Al-Barhamtoshy, K. M. Jambi, M. A. Rashwan, and S. M. Abdou, "An Arabic Manuscript Regions Detection, Recognition and Its Applications for OCRing," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 1, Feb. 2023, doi: 10.1145/3532609.

[46] B. Y. Al-Khuraym and M. M. Ben Ismail, "Arabic Sign Language Recognition using Lightweight CNN-based Architecture," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 319–328, 2022, doi: 10.14569/IJACSA.2022.0130438.

[47] B. Svendsen and S. Kadry, "Comparative Analysis of Image Classification Models for Norwegian Sign Language Recognition," *Technologies (Basel)*, vol. 11, no. 4, Aug. 2023, doi: 10.3390/technologies11040099.

[48] I. Rodríguez-Moreno, J. M. Martínez-Otzeta, I. Goienetxea, and B. Sierra, "Sign language recognition by means of common spatial patterns: An analysis," *PLoS One*, vol. 17, no. 10 October, Oct. 2022, doi: 10.1371/journal.pone.0276941.

[49] T. J. Sánchez-Vicinaiz, E. Camacho-Pérez, A. A. Castillo-Atoche, M. Cruz-Fernandez, J. R. García-Martínez, and J. Rodríguez-Reséndiz, "MediaPipe Frame and Convolutional Neural Networks-Based Fingerspelling Detection in Mexican Sign Language," *Technologies (Basel)*, vol. 12, no. 8, Aug. 2024, doi: 10.3390/technologies12080124.

[50] A. Kinanti and D. Maulana, "Convolutional Neural Network Implementation in BISINDO Alphabet Sign Language Recognition System Using Flask," *International Journal of New Media Technology)*, vol. 11, no. 1, p. 16, 2024.

[51] Z. Alsaadi, E. Alshamani, M. Alrehaili, A. A. D. Alrashdi, S. Albelwi, and A. O. Elfaki, "A Real Time Arabic Sign Language Alphabets (ArSLA) Recognition Model Using Deep Learning Architecture," *Computers*, vol. 11, no. 5, May 2022, doi: 10.3390/computers11050078.

[52] F. S. Alamri, A. Rehman, S. B. Abdullahi, and T. Saba, "Intelligent Real-Life Key-Pixel Image Detection System for Early Arabic Sign Language Learners," *PeerJ Comput Sci*, vol. 10, 2024, doi: 10.7717/PEERJ-CS.2063.

[53] R. R. Agustin, H. Maulana, and E. P. Mandyartha, "Detection of Actions BISINDO (Indonesian Sign Language) into Text-to-Speech using Long Short-Term Memory with MediaPipe Holistics," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp.

1051–1061, Nov. 2023, doi: 10.52436/1.jutif.2024.5.4.1492.

[54] R. Cahuantzi, X. Chen, and S. Güttel, "A Comparison of LSTM and GRU Networks for Learning Symbolic Sequences," in *Intelligent Computing*, K. Arai, Ed., Cham: Springer Nature Switzerland, 2023, pp. 771–785.

[55] U. Farooq, M. S. Mohd Rahim, and A. Abid, "A multi-stack RNN-based neural machine translation model for English to Pakistan sign language translation," *Neural Comput Appl*, vol. 35, no. 18, pp. 13225–13238, Jun. 2023, doi: 10.1007/s00521-023-08424-0.

[56] Noura Alshareef, Rema Abobake, and Asma Abd Aljalil, "Arabic Sign Language Recognition in Real Time Using Transfer Deep Learning," *AlQalam Journal of Medical and Applied Sciences*, pp. 730–739, Sep. 2024, doi: 10.54361/ajmas.247338.

[57] A. M. J. AL Moustafa *et al.*, "Integrated Mediapipe with a CNN Model for Arabic Sign Language Recognition," *Journal of Electrical and Computer Engineering*, vol. 2023, pp. 1–15, Aug. 2023, doi: 10.1155/2023/8870750.

[58] M. Mustafa, "Retraction Note to: A study on Arabic sign language recognition for differently abled using advanced machine learning classifiers," *J Ambient Intell Humaniz Comput*, vol. 14, no. 1, p. 381, 2023, doi: 10.1007/s12652-022-04142-y.

[59] R. Subandi, . H., and A. Yudhana, "Pneumonia Medical Image Classification Using Convolution Neural Network Model AlexNet and GoogleNet," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp. 1675–1684, Oct. 2024, doi: 10.12785/ijcds/1601124.

[60] X. Wu *et al.*, "CTransCNN: Combining transformer and CNN in multilabel medical image classification," *Knowl Based Syst*, vol. 281, Dec. 2023, doi: 10.1016/j.knosys.2023.111030.

[61] S. Saifullah *et al.*, "Nondestructive Chicken Egg Fertility Detection Using CNN-Transfer Learning Algorithms," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 3, pp. 854–871, 2023, doi: 10.26555/jiteki.v9i3.26722.

[62] D. O. Melinte and L. Vladareanu, "Facial Expressions Recognition for Human–Robot Interaction Using Deep Convolutional Neural Networks with Rectified Adam Optimizer," *Sensors*, vol. 20, no. 8, 2020, doi: 10.3390/s20082393.

[63] P. Singh, M. Kansal, R. Singh, S. Kumar, and C. Sen, "A Hybrid Approach based on Haar Cascade, Softmax, and CNN for Human Face Recognition," *J Sci Ind Res (India)*, vol. 83, no. 4, pp. 414–423, Apr. 2024, doi: 10.56042/jsir.v83i4.3167.

[64] M. T. Hoang, B. Yuen, X. Dong, T. Lu, R. Westendorp, and K. Reddy, "Recurrent Neural Networks for Accurate RSSI Indoor Localization," *IEEE Internet Things J*, vol. 6, no. 6, pp. 10639–10651, 2019, doi: 10.1109/JIOT.2019.2940368.

[65] G. H. Samaan *et al.*, "MediaPipe's Landmarks with RNN for Dynamic Sign Language Recognition," *Electronics (Switzerland)*, vol. 11, no. 19, Oct. 2022, doi: 10.3390/electronics11193228.

[66] K. K. Podder *et al.*, "Signer-Independent Arabic Sign Language Recognition System Using Deep Learning Model," *Sensors*, vol. 23, no. 16, Aug. 2023, doi: 10.3390/s23167156.

[67] W. N. Waluyo, R. Rizal Isnanto, and Adian Fatchur Rochim, "Comparison of Mycobacterium Tuberculosis Image Detection Accuracy Using CNN and Combination CNN-KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 1, pp. 80–87, Feb. 2023, doi: 10.29207/resti.v7i1.4626.

[68] C. C. Wang, C. Te Chiu, and J. Y. Chang, "EfficientNet-eLite: Extremely Lightweight and Efficient CNN Models for Edge Devices by Network Candidate Search," *J Signal Process Syst*, vol. 95, no. 5, pp. 657–669, May 2023, doi: 10.1007/s11265-022-01808-w.

[69] D. Tresnawati, R. Algani, and S. Mubaraq, "The Introduction of Hijaiyah Letters in Sign Languages using Augmented Reality Technology," *Jurnal Teknik Informatika (Jutif)*, vol. 3, no. 4, pp. 907–913, Aug. 2022, doi: 10.20884/1.jutif.2022.3.4.368.

[70] M. M. Taye, "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions," Mar. 01, 2023, *MDPI*. doi: 10.3390/computation11030052.