With around 1 % of the population of the Republic of Kazakhstan being affected by hearing disabilities, Kazakh Sign Language holds great importance as a means of communication between citizens of the state. The limitations of tools for Kazakh Sign Language (KSL) create significant challenges for people with hearing impairments in education, employment, and daily interactions. This research addresses these challenges through the development of an automated recognition system for Kazakh Sign Language gestures, aiming to enhance accessibility and inclusivity of communication using artificial intelligence. The approach employs advanced machine learning techniques, including Convolutional Neural Networks (CNNs) for recognizing spatial gesture patterns and Recurrent Neural Networks (RNNs) for processing temporal sequences. By combining these methods, the system recognizes both hand gestures and facial expressions, providing a dual-stream model that surpasses single-stream gesture recognition systems focused solely on hand movements. A dedicated dataset was created using Mediapipe Holistic, an open-source tool that identifies 543 landmarks across hands, faces, and poses, effectively capturing the multifaceted nature of sign language. The findings showed that the hybrid model significantly outperformed standalone CNN and RNN models, achieving up to 96 % accuracy. This demonstrates that integrating facial expressions with hand gestures greatly enhances the precision of sign language recognition. This system holds immense potential to improve inclusivity and accessibility in various settings across the Republic of Kazakhstan by facilitating communication for hearing-impaired individuals, paving the way for expanded research and application in other sign languages

Keywords: dynamic gesture recognition, hybrid neural network, Kazakh Sign Language, facial expressions

UDC 004.89
DOI: 10.15587/1729-4061.2025.315834

DEVELOPMENT OF A HYBRID CNN-RNN MODEL FOR ENHANCED RECOGNITION OF DYNAMIC GESTURES IN KAZAKH SIGN LANGUAGE

Aigerim Aitim
Master of Technical Sciences, Assistant-Professor*
Dariga Sattarkhuzhayeva
Corresponding author
Bachelor of Information Communication Technologies*
E-mail: sattarkhuzhaevadariga@gmail.com
Aisulu Khairullayeva
Bachelor of Information Communication Technologies*
*Department of Information Systems
International Information Technology University
Manas str., 34/1, Almaty, Republic of Kazakhstan, 050020

Received 22.11.2024
Received in revised form 20.02.2025
Accepted date 10.03.2025
Published date 22.04.2025

How to Cite: *Aitim, A., Sattarkhuzhayeva, D., Khairullayeva, A. (2025). Development of a hybrid CNN-RNN model for enhanced recognition of dynamic gestures in Kazakh Sign Language. Eastern-European Journal of Enterprise Technologies, 2 (2 (134)), 58–67. https://doi.org/10.15587/1729-4061.2025.315834*

1. Introduction

The rapid development of science and technology has significantly contributed to solving various societal challenges, particularly those faced by individuals with disabilities. For people with hearing impairments, effective communication remains a vital but often complex aspect of daily life. Sign language, which involves the use of hand movements and facial expressions, serves as their primary mode of communication. However, communication barriers between hearing-impaired individuals and the rest of society frequently lead to social isolation [1].

In the Republic of Kazakhstan, there are 200,000 people with hearing impairments, which constitutes about 1 % of the country's 20 million population [2, 3]. These individuals encounter numerous challenges in education, employment, and social inclusion, primarily due to the limited availability of sign language interpretation services. Currently, the state provides only 60 hours of interpretation services annually per person [4], while several regions face a severe shortage of professional interpreters [5]. These issues underscore the need for innovative solutions that can either support or replace traditional interpretation services, improving accessibility and the overall quality of life for this group.

Advancements in artificial intelligence (AI), machine learning (ML), and computer vision have opened up new possibilities for addressing such challenges. Machine learning techniques, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have already demonstrated considerable success in recognizing sign languages through the analysis of hand gestures and facial expressions. Research in this field highlights the effectiveness of these technologies. For instance, studies have successfully developed recognition systems for Assamese Sign Language using MediaPipe combined with deep learning models [6]. Other efforts have applied machine learning techniques to recognize the Amharic alphabet [7]. Indian Sign Language recognition has utilized CNNs for dynamic gesture modeling and Support Vector Machines (SVMs) for image analysis [8, 9]. Additionally, Arabic Sign Language [10] and Indonesian Sign Language [11] have been the focus of similar research, incorporating hand gesture recognition to improve communication tools for hearing-impaired individuals.

Nevertheless, many existing systems primarily concentrate on recognizing hand gestures, neglecting the significant role that facial expressions play in effective sign language communication. This research aims to address this gap by developing a dual-stream model that integrates CNN and RNN technologies to recognize the Kazakh Sign Language (KSL). By utilizing Mediapipe Holistic, an open-source framework capable of detecting 543 key points, including 33 body poses, 21 landmarks for each hand, and 468 facial landmarks [12], the proposed system will capture both hand and facial gestures. Such a tool could provide

an innovative solution to improve communication for people with hearing impairments in the Republic of Kazakhstan.

Kazakh Sign Language differs from many other sign languages in its dynamic nature and its extensive use of facial expressions alongside hand gestures. These unique characteristics make it particularly challenging for conventional recognition systems. Therefore, the exploration of methods that can effectively recognize the Kazakh Sign Language is highly relevant. Such research not only addresses a critical societal need but also contributes to the broader field of sign language recognition technologies, fostering inclusion and enhancing the quality of life for individuals with hearing impairments in the Republic of Kazakhstan.

2. Literature review and problem statement

Several studies have been conducted by researchers and organizations to address challenges in sign language recognition, leveraging advancements in machine learning and computer vision technologies. Despite these efforts, Kazakh Sign Language (KSL) remains an underexplored area due to its unique characteristics, such as dynamic gestures and the use of facial expressions. This analysis explores existing studies, highlighting their contributions, limitations, and relevance to the recognition of KSL.

The paper [13] proposed a vision-based system for recognizing gestures in Thai Sign Language (TSL). Their method employed Zernike moments to reduce computation time when analyzing single-handed and double-handed gestures. This approach is particularly effective for real-time applications, such as mobile platforms. However, its adaptability to dynamic languages like KSL remains unclear, especially given the importance of integrating gestures with facial expressions. Developing algorithms that incorporate both elements is a potential solution.

In this study [14], researchers combined surface electromyography (sEMG) signals with Temporal Convolutional Networks (TCN) and Long Short-Term Memory (LSTM) models. This multi-modal framework significantly improved the recognition accuracy of complex gestures. For KSL, however, the lack of annotated datasets for hand gestures and facial expressions limits the direct application of such techniques. Expanding these datasets and integrating facial landmarks into the framework could enhance its relevance.

A significant contribution was made in this study [15] with the creation of a large-scale dataset for Greek Sign Language (GSL), which demonstrated improved recognition accuracy through deep learning models. The absence of a similar dataset for KSL remains a barrier to robust model training. Collaborative efforts to build a comprehensive KSL dataset could address this issue, enabling better performance in recognition systems.

The fusion of handcrafted skeleton-based features with pixel-based deep learning features for Japanese Sign Language (JSL) recognition was explored in this study [16]. This approach enhanced robustness against variability in gestures and signer styles. For KSL, employing similar techniques alongside tools like Mediapipe Holistic, which can simultaneously capture hand and facial landmarks, could offer a holistic solution to recognition challenges.

In this study [17], a hybrid-metaheuristic algorithm for feature selection achieved an impressive 98 % accuracy rate in gesture recognition. The optimized feature sets proved essential in enhancing machine learning model performance. However, applying such optimization methods to KSL, with its diverse and complex gestures, may require tailored approaches for multi-modal datasets.

In paper [18], the authors conducted research on Kazakh dactylic sign language recognition using machine learning methods like Random Forest and Support Vector Machines (SVM). Their dataset of 5,000 images per gesture yielded respectable accuracy. However, their models did not support the recognition of sequential and dynamic gestures in KSL, which necessitates the use of advanced methods such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

In the study [19], an ensemble method combining ResNet-50 and VGG-19 architectures was applied to KSL gesture recognition, achieving a recognition accuracy of 95.7 %. While this method demonstrated effectiveness in static gesture recognition, it did not incorporate facial expressions, which are critical for accurately interpreting the full meaning of gestures in KSL. Incorporating multimodal inputs, such as hand gestures and facial expressions, could significantly improve recognition accuracy.

All this allows to argue that it is appropriate to conduct a study devoted to developing advanced methods for the recognition of Kazakh Sign Language (KSL). Despite substantial progress in sign language recognition, critical gaps remain in addressing the unique characteristics of KSL, including its dynamic gestures and reliance on facial expressions. Addressing these challenges requires the development of specific methods, including the creation of large-scale annotated datasets and the application of dual-stream methods that utilize Convolutional Neural Networks (CNNs) for static gesture recognition and Recurrent Neural Networks (RNNs) for processing sequential data. Such methods can significantly enhance the recognition of complex gestures and facial expressions in KSL, ultimately enhancing communication tools for individuals with hearing impairments in the Republic of Kazakhstan.

3. The aim and objectives of the study

The aim of this study is to develop a hybrid CNN-RNN model to accurately recognize dynamic and facial gestures in Kazakh Sign Language (KSL). The research focuses on creating a practical translation system that enhances communication for individuals with hearing impairments in the Republic of Kazakhstan, promoting social inclusion and bridging communication gaps through advanced gesture recognition technology.

To accomplish this aim, the next objectives were set:

– to create a comprehensive dataset with precise coordinates of facial and hand landmarks, covering both static and dynamic gestures in KSL to ensure diverse and accurate gesture representation;

– to design a hybrid architecture the Gesture Recognition System;

– to evaluate the performance of the hybrid model in recognizing dynamic and static gestures of KSL, comparing its accuracy with traditional single CNN-based models;

– to apply the developed model in practical tools, such as real-time translation systems, to improve communication for individuals with hearing impairments.

4. Methods and materials

4. 1. Object and hypothesis of the study

The object of this study is the process of recognizing dynamic and facial gestures in Kazakh Sign Language (KSL) using a

hybrid CNN-RNN model, focusing on developing a robust method that integrates both manual and facial gesture recognition to enhance communication for individuals with hearing impairments.

The main hypothesis of the study is that the hybrid CNN-RNN model will demonstrate significantly higher accuracy in recognizing dynamic and facial gestures of KSL compared to traditional single-stream CNN models.

Assumptions made in the study are that all gestures are performed in a controlled environment with consistent lighting and a neutral background, that the dataset accurately represents the diversity of gestures and facial expressions in KSL, and that the hybrid model will maintain performance consistency in varied real-world scenarios.

Simplifications adopted in the study are the recognition process is simplified by limiting the dataset to a predefined set of gestures and facial expressions, restricting dynamic gestures to specific motion patterns to reduce model complexity, and minimizing background noise and non-relevant movements during data collection to improve model accuracy.

This structured method ensures that the study not only validates the effectiveness of the proposed hybrid model but also explores its potential for real-world applications, contributing to the broader goal of enhancing social inclusion for the hearing-impaired community in the Republic of Kazakhstan.

4. 2. Theoretical methods

The recognition system was based on a hybrid architecture combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs were used to analyze spatial features such as hand shapes and facial expressions, essential for static gesture recognition. RNNs captured temporal dependencies in sequential frames, making them suitable for dynamic gestures. The hybrid model was selected for its ability to handle multimodal data effectively, as demonstrated in similar sign language recognition studies [6, 7].

Additionally, Mediapipe Holistic was employed for extracting multimodal features, including 21 hand landmarks, 468 facial points, and 33 body posture points [12].

4. 3. Data collection

The data collection process was carefully structured to ensure high-quality and representative inputs for training the models:

Video recording:

– gestures were recorded using OpenCV, utilizing Full HD cameras under controlled lighting conditions to reduce noise and enhance clarity;

– a neutral background was used to eliminate distractions;

– the dataset included a diverse set of both static gestures and dynamic gestures;

– an automated data collection algorithm was impleł mented to streamline the capture of video data and preprocessing of frames for analysis.

Feature extraction:

– mediapipe Holistic was employed to extract spatial and temporal features from the recorded videos;

– the extracted features included hand shapes, facial expressions, and body postures, ensuring a comprehensive representation of gestures.

This structured data collection ensured that the dataset accurately represented the complexity and diversity of KSL gestures.

4. 4. Data preprocessing

The collected data underwent meticulous preprocessing to ensure its readiness for model training. The extracted features were stored as NumPy arrays, where each array contained the (x, y, z) coordinates of all detected landmarks. Missing values were replaced with zeros to maintain uniformity across the dataset. To enhance model efficiency, data normalization was performed, standardizing the input values and ensuring compatibility with machine learning algorithms.

The dataset was then divided into two subsets: 80 % allocated for training to facilitate pattern learning and 20 % reserved for testing, providing a robust framework for evaluating model performance.

This comprehensive preprocessing pipeline ensured the data's consistency, clarity, and suitability for training and validation.

4. 5. Model development

Three distinct models were developed to process the preprocessed data:

1. Convolutional Neural Networks (CNNs).

CNNs were developed to analyze spatial features, focusing on static gestures such as hand shapes and facial expressions. These models excel at recognizing stationary gestures in Kazakh Sign Language (KSL) by extracting patterns from individual video frames. Their proven ability to detect spatial features made them a vital component of the system [8, 10].

2. Recurrent Neural Networks (RNNs).

RNNs were utilized for the analysis of sequential data, capturing the temporal relationships required for interpreting dynamic gestures. This approach was well-suited for recognizing movements that spanned multiple frames, such as directional gestures, by processing frame-to-frame dependencies. The inclusion of RNNs ensured the system could accurately track motion patterns [7, 11].

3. Hybrid CNN-RNN Model.

A hybrid model was implemented to integrate the strengths of CNNs and RNNs, allowing simultaneous processing of static and dynamic gestures. This combined approach ensured effective recognition of complex KSL gestures, leveraging CNNs for spatial feature extraction and RNNs for temporal sequence analysis.

The models were implemented and configured to address the unique requirements of KSL gestures.

4. 6. Validation

The developed models underwent a rigorous validation process to ensure their reliability and effectiveness. Each model was trained over 1,000 epochs, enabling the networks to learn and generalize gesture patterns effectively. The performance was evaluated by comparing the predicted gestures with ground truth labels, using metrics such as categorical accuracy and recognition rates to assess their ability to recognize both static and dynamic gestures.

Categorical accuracy measures how well a model predicts the correct class in a dataset where each class is represented by a unique label (e. g., letters like A, B, V). For each record, the model predicts probabilities or scores for all classes, and the class with the highest value is selected as the predicted label. This predicted label is then compared with the actual label. If they match, it is counted as correct; otherwise, it is incorrect. The accuracy is calculated as the percentage of correct predictions out of the total records. For example, if there are 100 samples and the model correctly predicts 92 of them, the accuracy is 92 %. Categorical accuracy is a simple and

effective way to measure performance, especially in tasks like recognizing letters, where each class is distinct, and the model aims to match the predicted letter with the actual one:

$$\text{Categorical accuracy} = \\ = \frac{1}{N}\sum_{i=1}^{N} 1\left(\arg\max\left(y_{pred}^{(i)}\right) = \arg\max\left(y_{true}^{(i)}\right)\right), \quad (1)$$

where N is total number of frames, $y_{pred}^{(i)}$ is predicted class for i-th frame, $y_{true}^{(i)}$ is actual class for i-th frame. Arg max function is used to find maximum value from found values and 1(•) is indicator function that results in 1 if condition inside is true and in 0 if not.

Extensive testing of the hybrid CNN-RNN model demonstrated its capability to handle the multimodal complexities of Kazakh Sign Language (KSL) gestures. This validation process confirmed the robustness and adaptability of the hybrid model, underscoring its potential for real-world applications in KSL translation systems.

5. Results of hybrid CNN-RNN model application for recognizing KSL gestures

5. 1. Creation of a comprehensive dataset for Kaf zakh Sign Language gestures

To enable accurate recognition of Kazakh Sign Language (KSL) gestures, a robust dataset was developed. This dataset integrates hand and facial landmarks to ensure a precise representation of gestures.

To collect the data, OpenCV was used to process each frame of the video, where algorithms were then applied to extract key features as can be seen in Fig. 1. A major task at this stage was ensuring the clarity of the video material and accuracy in capturing the movements to guarantee the correctness of the data passed to the next stage.

After collecting data using OpenCV, Mediapipe Holistic was used to analyze the movements. This solution captures 543 key points covering both the hands and face, making it an ideal tool for analyzing Kazakh Sign Language gestures. Mediapipe Holistic extracts 21 key points on each hand, allowing precise tracking of finger and palm positions. These data are particularly important for static gestures like letters and numbers, where the shape of the hands is a key element. For facial analysis, 468 key points were captured, and 33 key points were recorded for body posture. Facial expressions play an important role in Kazakh Sign Language, and these data provide accurate recognition of emotional and contextual elements of gestures.

The built-in Mediapipe Landmark Drawing Tool was used for visualizing the data, displaying key points on the face and hands directly on the video as displayed in Fig. 2. This allowed real-time tracking of gesture and facial expression recognition accuracy, making it easier to identify and fix errors, which improved the data quality for training the models.

After extracting key points from each video frame, they were saved as NumPy arrays, with each array containing the coordinates (*x*, *y*, *z*) of all detected points on the hands, face, and body in Fig. 3. If any point was not detected, its coordinates were filled with zeros. These arrays were combined into a general dataset where each frame contains information about the position of all key points in space. For training purposes, the data was normalized and split into training and testing sets.

```
with mp_holistic.Holistic(min_detection_confidence=0.6, min_tracking_confidence=0.6) as holistic_model:
    for label in labels:
        if exit_program:
            break
        for sample_id in range(samples_per_label):
            if exit_program:
                break
            for frame_index in range(frames_per_sample):
                ret, frame = camera_stream.read()


                processed_frame, detections = process_frame_with_model(frame, holistic_model)
                render_custom_landmarks(processed_frame, detections)

                if frame_index == 0:
                    cv2.putText(processed_frame, 'STARTING COLLECTION', (50, 150),
                                cv2.FONT_HERSHEY_SIMPLEX, 1, (0, 255, 0), 3, cv2.LINE_AA)
                    cv2.putText(processed_frame, f'Collecting: {label}, Sample: {sample_id}', (10, 30),
                                cv2.FONT_HERSHEY_SIMPLEX, 0.7, (0, 0, 255), 2, cv2.LINE_AA)
                    cv2.imshow(window_name, processed_frame)
                    cv2.waitKey(200)  # Pause for a moment to show "STARTING COLLECTION"
                else:
                    # Display collection progress
                    collection_message = f'Collecting: {label}, Sample: {sample_id}, Frame: {frame_index}'
                    cv2.putText(processed_frame, collection_message, (10, 30),
                                cv2.FONT_HERSHEY_SIMPLEX, 0.7, (255, 255, 255), 2, cv2.LINE_AA)
                    cv2.imshow(window_name, processed_frame)


                # Save the detected keypoints
                keypoints = extract_keypoints_from_results(detections)
                save_path = os.path.join(OUTPUT_FOLDER, label, str(sample_id), f'{frame_index}.npy')
                np.save(save_path, keypoints)

                # Check for user input to exit
                if cv2.waitKey(10) & 0xFF == ord('q'):
                    exit_program = True
                    break
camera_stream.release()
cv2.destroyAllWindows()
```

Fig. 1. Algorithm of dataset collection

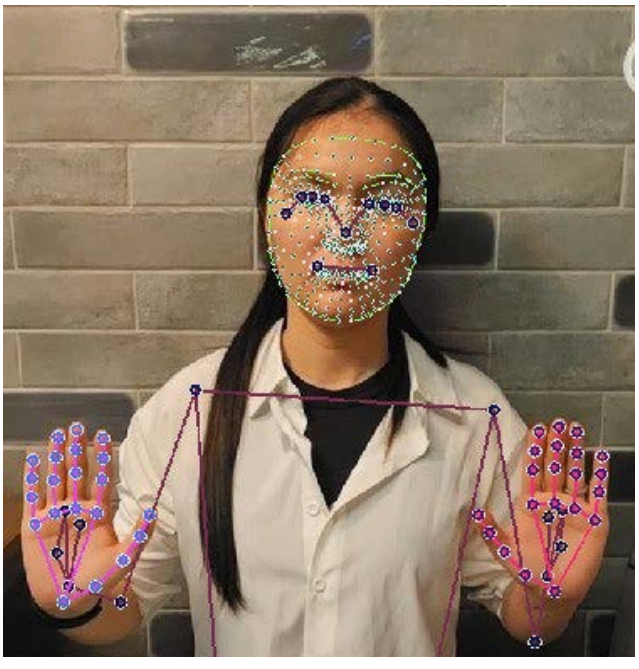


Fig. 2. Displaying face and hand landmarks using Mediapipe Holistic and Mediapipe Drawing Tools

```
>>> results.face_landmarks
...
landmark {
  x: 0.482445
  y: 0.52115947
  z: -0.0261088572
}
landmark {
  x: 0.482070386
  y: 0.474325955
  z: -0.0451906361
}
landmark {
  x: 0.482135177
  y: 0.488987446
  z: -0.0243302248
}
landmark {
  x: 0.472986609
  y: 0.423679054
  z: -0.0325737186
}
```

a

```
>>> results.right_hand_landmarks
...
landmark {
  x: 0.214127183
  y: 0.848286927
  z: 4.26226961e-007
}
landmark {
  x: 0.253028572
  y: 0.819111705
  z: -0.0234541688
}
landmark {
  x: 0.288173378
  y: 0.78141588
  z: -0.0392999537
}
landmark {
  x: 0.321514934
  y: 0.76109159
  z: -0.0542295501
}
```

b

```
>>> results.left_hand_landmarks
...
landmark {
  x: 0.737200916
  y: 0.776185036
  z: 3.64397209e-007
}
landmark {
  x: 0.694661736
  y: 0.759467125
  z: -0.0184820145
}
landmark {
  x: 0.654934
  y: 0.731037378
  z: -0.0306072254
}
landmark {
  x: 0.620742619
  y: 0.724801898
  z: -0.0418048
}
```

c

Fig. 3. Coordinates (*x, y, z*) of positions of landmarks:
a — face landmarks; *b* — right hand landmarks; *c* — left hand landmarks

This dataset provides a detailed and diverse foundation for training machine learning models, ensuring precision and robustness in gesture recognition.

5. 2. Development the hybrid architecture

CNNs were used for spatial analysis, focusing on static gestures such as hand shapes and facial expressions, while RNNs captured the sequential nature of dynamic gestures through temporal analysis. The integration of these two components into a hybrid architecture, as depicted in Fig. 4, allowed the system to process both spatial and temporal features simultaneously.

The overall structure of the gesture recognition system is illustrated in Fig. 5, showcasing the stages from video input, feature extraction using Mediapipe, preprocessing, and ultimately classification using the hybrid model.

This robust architecture enabled accurate recognition of both static and dynamic gestures, addressing the KSL complexity.

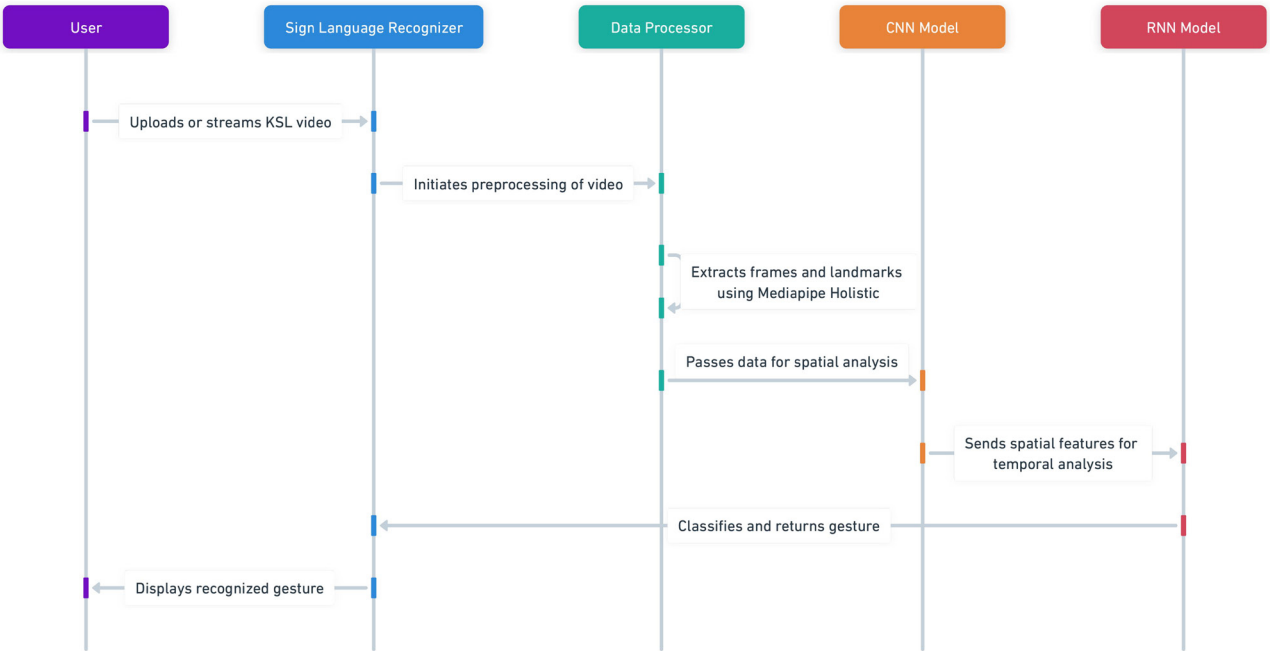


Fig. 4. Hybrid CNN-RNN gesture recognition process

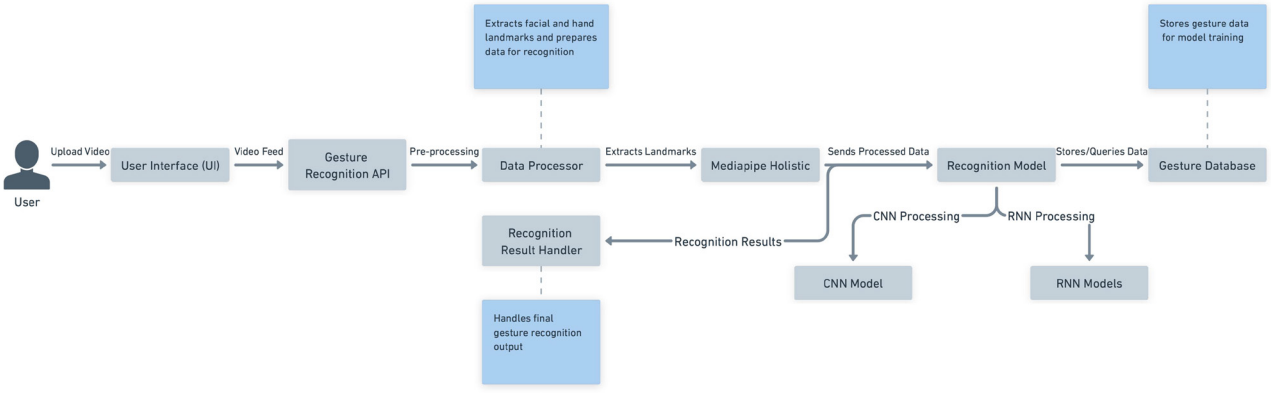


Fig. 5. Structure of the gesture recognition system

5. 3. Performance evaluation of machine learning models

The performance of the three machine learning models in recognizing Kazakh Sign Language (KSL) gestures was evaluated through comparison of the first and the last epoch of the training. The results indicate significant improvements in accuracy as the training progressed, highlighting the effectiveness of the hybrid model.

In the first epoch, as shown in Table 1, the accuracy rates for recognizing three letters of KSL were relatively low across all models. The CNN resulted in 25 % accuracy for the recogu nition of letter "V," 28 % of letter "O," and 29 % of letter "P." The RNN performed better, with 38 %, 37 %, and 37 % preR cision of recognition, respectively. The hybrid model, which combines both CNN and RNN, demonstrated a marked improvement, reaching 47 % for "V," 43 % for "O," and 40 % for "P." These initial results underscore the potential of the hybrid approach in recognizing KSL gestures.

By the end of the training period at the epoch 1000, the performance of the models improved significantly, as depicted in Table 2. The CNN model achieved accuracies of 75 % for "V," 74 % for "O," and 71 % for "P." The RNN model showed moderate improvement with Categorical accuracy rates of 59 %, 61 %, and 60 %, respectively. In contrast, the hybrid model exhibited exceptional results, with accuracies soaring to 96 % for "V," 93 % for "O," and 91 % for "P." These final results demonstrate the effectiveness of using a hybrid approach for KSL recognition, as it not only improves categorical accuracy but also captures the nuances of both gestures and facial expressions over time.

Table 1

Results in epoch 1 out of 1000 for 3 letters of Kazakh Sign Language in three different approaches

Machine learning models	V	O	P
CNN	25 %	28 %	29 %
RNN	38 %	37 %	37 %
Hybrid of CNN and RNN	47 %	43 %	40 %

To further illustrate the dataset used in this study, Fig. 6 provides a frame showcasing the letter "V" in Kazakh Sign Language. This visual representation emphasizes the importance of incorporating diverse data points, including various

hand positions and facial expressions, to train the models effectively. Overall, the findings suggest that the hybrid model significantly outperforms individual models, particularly in the context of recognizing KSL gestures. This validates the hypothesis that combining CNN and RNN architectures can enhance the recognition categorical accuracy of sign language, making it a promising approach for future research and applications in KSL translation.

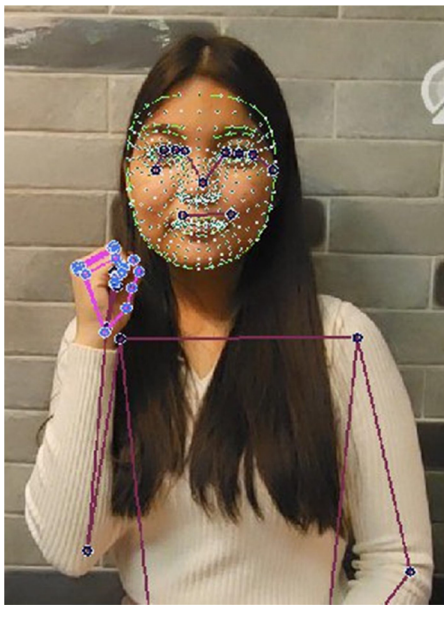


Fig. 6. Frame from dataset where letter V in Kazakh Sign Language

Table 2

Results in epoch 1000 out of 1000 for 3 letters of Kazakh Sign Language in three different approaches

Machine learning models	V	O	P
CNN	75 %	74 %	71 %
RNN	59 %	61 %	60 %
Hybrid of CNN and RNN	96 %	93 %	91 %

These results demonstrate the hybrid model's superiority in handling both static and dynamic gestures, validating its multimodal recognition capabilities.

5. 4. Practical implementations and future applications

The system developed for recognizing hand and facial gestures in Kazakh Sign Language (KSL) has the potential to make a significant difference in the lives of people with hearing disabilities. One of the most promising outcomes is the creation of a mobile application that translates KSL gestures into spoken language in real time. This app could help reduce the isolation often experienced by hearing-impaired individuals, enabling them to communicate more easily with others. By combining gesture recognition with facial expression analysis, the app ensures precise translations that capture not just the gestures but also their meaning and context, making it an essential tool for everyday communication.

The app is designed to be user-friendly and accessible to people of all ages and technical backgrounds. It will include features like:

– a gesture library: users can explore and learn KSL gestures at their own pace;

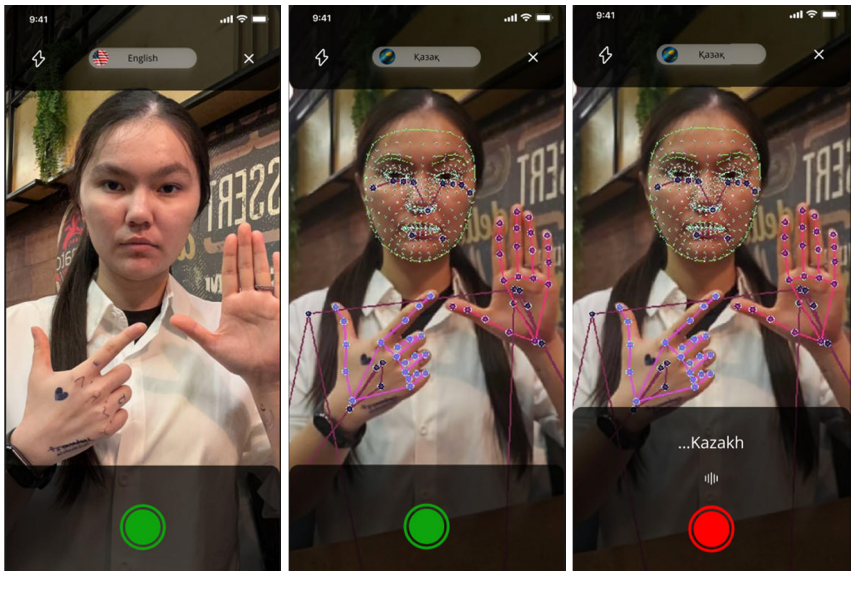
– real-time conversation mode: this mode allows instant translation of gestures into speech, making live conversations seamless;

– customization options: to make the app accessible to everyone, users can personalize settings such as text size or background colors.

The design of the application is shown in Fig. 7, highlighting: how the app looks before recognition starts; the detection of key points on the hands and face during gesture recognition; the final step where the gesture is translated into text or spoken language.

Beyond daily communication, this app could be a game-changer in sectors like education, healthcare, and customer service. In schools, it could help translate lessons into KSL, ensuring that students with hearing impairments have equal opportunities to learn. In healthcare, it could bridge communication gaps between patients and medical professionals, leading to better care. In customer service, businesses could use the app to interact more effectively with hearing-impaired customers.

This application isn't just a communication tool – it's a step toward a more inclusive society. By making interactions more accessible and recognizing the needs of the hearing-impaired community, it has the potential to foster understanding, reduce barriers, and create meaningful connections.



a *b* *c*

Fig. 7. The design of application: *a* – before recognition; *b* – highlighting key points; *c* – translation

6. Discussion of results of comparison of three machine learning models

The results obtained from this study highlight the significant advantages of the hybrid CNN-RNN model in recognizing

Kazakh Sign Language (KSL) gestures, as shown in Table 2. The model achieved accuracy rates of 96 %, 93 %, and 91 % for the letters "V," "O," and "P," respectively, by the 1000th epoch. These results underscore the effectiveness of combining CNNs for spatial analysis and RNNs for temporal sequence processing, as shown in Fig. 4, 5. This combination allows for more effective recognition of both static and dynamic gestures, addressing the limitations of standalone CNN and RNN models. The robustness of these results can also be attributed to the detailed dataset developed, with key point coordinates extracted using Mediapipe Holistic, as depicted in Fig. 3.

What makes the hybrid model particularly effective is its ability to balance high accuracy and computational efficiency. Advanced systems like the C3D-BiLSTM model for American Sign Language (ASL) have achieved slightly higher accuracy rates, such as 98.91 % [20]. However, these systems rely on significant computational resources, making them unsuitable for real-time or mobile applications. In contrast, the hybrid model proposed in this study achieves comparable accuracy while remaining lightweight enough for deployment in practical scenarios, including mobile devices.

The capability of the hybrid model to handle complex gestures is another notable feature. Existing systems for Indian Sign Language (ISL), for instance, perform well for simpler gestures but are limited when it comes to recognizing more intricate gestures involving multiple hands or the integration of facial expressions [8, 21, 22]. This limt itation is addressed by the hybrid model, which combines hand landmarks with facial expressions to provide a more comprehensive understanding of gestures. This broader recognition capability is evident in the model's high performance, as illustrated by the consistent improvement in accuracy from the first epoch (Table 1) to the final epoch (Table 2).

Despite its promising results, the study has certain limitations that should be addressed. The dataset used in this research, while detailed, includes only a specific range of KSL gestures. This restricts the system's ability to generalize to less common gestures or variations influenced by regional or individual differences. Expanding the dataset to incorporate more diverse gestures and contexts is a critical next step [5, 19]. Furthermore, the system's peru formance under controlled conditions has been validated, but real-world factors such as varying lighting, cluttered backgrounds, and overlapping gestures may pose challenges [6, 18]. These environmental sensitivities highlight the need for additional testing in diverse, uncontrolled settings.

Another limitation is the reliance on Mediapipe Holistic for feature extraction [12]. While this tool provides accurate and detailed landmarks, dependence on an external framework could limit adaptability if alternative tools or proprietary feature extraction methods are required. Developing a custom feature extraction pipeline tailored specifically for KSL recognition could mitigate this reliance and enhance system flexibility [18].

In addition to limitations, the study also has certain disadvantages. One key drawback is that the system requires preprocessing of data in controlled environments, which may not always be feasible in practical applications. To overcome this, future research could explore adaptive preprocessing techniques and the integration of noise reduction methods to improve system robustness in real-world scenarios [12].

Looking forward, the hybrid model offers several opportunities for further development. A practical application of this system would be a mobile application for real-time KSL translation, as envisioned in Fig. 7, which could bridge communication gaps in everyday life, education, and healthcare [23]. However, deploying the system on low-power mobile devices will require optimization of the model architecture to maintain its performance without increasing computational demands. Expanding the system to recognize gestures from other sign languages or incorporating multilingual capabilities could significantly broaden its scope and impact [21, 24–26].

Challenges in the development of this system may include managing the mathematical and algorithmic complexity of optimizing the hybrid architecture for resource-constrained devices, as well as the logistical difficulties of collecting and labeling a more diverse dataset. Real-world testing with diverse user groups and in varied environments will also be essential to validate the model's effectiveness and identify any further adjustments needed for widespread adoption [1, 2, 4].

In conclusion, while there are limitations and challenges to address, the results of this study demonstrate the hybrid CNN-RNN model's strong potential for advancing KSL recognition. Its balance of accuracy, efficiency, and adaptability positions it as a viable solution for improving communication accessibility for the hearing-impaired community in the Republic of Kazakhstan and beyond. Future efforts should focus on expanding the dataset, optimizing the model for mobile use, and testing it in real-world settings to ensure its effectiveness and usability.

7. Conclusions

1. A detailed dataset for Kazakh Sign Language (KSL) was developed, capturing 543 key landmarks, including hand, facial, and body posture points. This dataset, featuring both static and dynamic gestures, proved sufficient for training machine learning models effectively. No overfitting was observed during training, confirming that the dataset size was appropriate without requiring additional augmentation.

2. The hybrid CNN-RNN model showed excellent results in recognizing KSL gestures, achieving accuracy rates of 96 % for "V," 93 % for "O," and 91 % for "P." These outcomes underline the strength of combining CNNs for analyzing static features with RNNs for processing gesture sequences. The model's ability to handle complex gestures involving hand movements and facial expressions makes it highly effective and versatile.

3. The use of Mediapipe Holistic for extracting hand and facial landmarks played a key role in the model's success, ensuring precision in feature detection. However, some gestures, particularly those with intricate or overlapping movements, could benefit from further refinement. Addressing these challenges will enhance the model's overall recognition accuracy.

4. The study establishes a practical basis for developing a mobile application for real-time KSL-to-speech translation. Such an app could significantly improve communication for people with hearing impairments in Kazakhstan,

enhancing accessibility in education, healthcare, and daily life. Practical recommendations include focusing on refining complex gesture recognition, integrating the model into mobile platforms, and conducting usability tests to validate the application in diverse real-world scenarios.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Funding

The study was performed without financial support.

Data availability

Data will be made available on reasonable request.

Use of artificial intelligence

The authors confirm that no artificial intelligence technologies were used in the creation of this work.

References

1. In Almaty, more than 10,000 people live with hearing disabilities (2022). Vecher.kz. Available at: https://vecher.kz/ru/article/v-almaty-projivaiut-bolee-10-tysiach-liudei-imeiushih-invalidnost-po-sluhu.html
2. People with hearing disabilities find it difficult to obtain a profession (2024). 24KZ. Available at: https://24.kz/ru/news/social/item/675566-lyudyam-s-invalidnostyu-po-slukhu-slozhno-poluchit-professiyu
3. On the demographic situation for January-September 2024 (2024). Statistical Committee of the Ministry of National Economy of the Republic of Kazakhstan. Available at: https://stat.gov.kz/ru/news/o-demograficheskoy-situatsii-za-yanvar-sentyabr-2024-goda/
4. Ibadullaeva, A. (2023). Guides from the World of Silence: How Sign Language Interpreters Work in Kazakhstan. Liter.kz. Available at: https://liter.kz/provodniki-iz-mira-tishiny-kak-rabotaiut-surdoperevodchiki-v-kazakhstane-1676962474/
5. Cheh, E. (2024). Acute shortage of hearing impairment educators in East Kazakhstan. Ustinka LIVE. Available at: https://ustinka.kz/vko/97381.html
6. Bora, J., Dehingia, S., Boruah, A., Chetia, A. A., Gogoi, D. (2023). Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning. Procedia Computer Science, 218, 1384–1393. https://doi.org/10.1016/j.procs.2023.01.117
7. Salau, A. O., Tamiru, N. K., Abeje, B. T. (2024). Derived Amharic alphabet sign language recognition using machine learning methods. Heliyon, 10 (19), e38265. https://doi.org/10.1016/j.heliyon.2024.e38265
8. Katoch, S., Singh, V., Tiwary, U. S. (2022). Indian Sign Language recognition system using SURF with SVM and CNN. Array, 14, 100141. https://doi.org/10.1016/j.array.2022.100141
9. Singh, D. K. (2021). 3D-CNN based Dynamic Gesture Recognition for Indian Sign Language Modeling. Procedia Computer Science, 189, 76–83. https://doi.org/10.1016/j.procs.2021.05.071
10. Ibrahim, N. B., Selim, M. M., Zayed, H. H. (2018). An Automatic Arabic Sign Language Recognition System (ArSLRS). Journal of King Saud University - Computer and Information Sciences, 30 (4), 470–477. https://doi.org/10.1016/j.jksuci.2017.09.007
11. Indra, D., Purnawansyah, Madenda, S., Wibowo, E. P. (2019). Indonesian Sign Language Recognition Based on Shape of Hand Gesture. Procedia Computer Science, 161, 74–81. https://doi.org/10.1016/j.procs.2019.11.101
12. MediaPipe Holistic – Simultaneous Face, Hand and Pose Prediction, on Device (2020). Google Research Blog. Available at: https://research.google/blog/mediapipe-holistic-simultaneous-face-hand-and-pose-prediction-on-device/
13. Chansri, C., Srinonchat, J. (2016). Hand Gesture Recognition for Thai Sign Language in Complex Background Using Fusion of Depth and Color Video. Procedia Computer Science, 86, 257–260. https://doi.org/10.1016/j.procs.2016.05.113
14. Shin, J., Miah, A. S. M., Konnai, S., Takahashi, I., Hirooka, K. (2024). Hand gesture recognition using sEMG signals with a multi-stream time-varying feature enhancement approach. Scientific Reports, 14 (1). https://doi.org/10.1038/s41598-024-72996-7
15. Papadimitriou, K., Sapountzaki, G., Vasilaki, K., Efthimiou, E., Fotinea, S.-E., Potamianos, G. (2024). A large corpus for the recognition of Greek Sign Language gestures. Computer Vision and Image Understanding, 249, 104212. https://doi.org/10.1016/j.cviu.2024.104212
16. Shin, J., Hasan, Md. A. M., Miah, A. S. M., Suzuki, K., Hirooka, K. (2024). Japanese Sign Language Recognition by Combining Joint Skeleton-Based Handcrafted and Pixel-Based Deep Learning Features with Machine Learning Classification. Computer Modeling in Engineering & Sciences, 139 (3), 2605–2625. https://doi.org/10.32604/cmes.2023.046334
17. Aitim, A., Satybaldiyeva, R. (2025). A comparison of Kazakh language processing models for improving semantic search results. Eastern-European Journal of Enterprise Technologies, 1 (2 (133)), 66–75. https://doi.org/10.15587/1729-4061.2025.315954
18. Kenshimov, C., Buribayev, Z., Amirgaliyev, Y., Ataniyazova, A., Aitimov, A. (2021). Sign language dactyl recognition based on machine learning algorithms. Eastern-European Journal of Enterprise Technologies, 4 (2 (112)), 58–72. https://doi.org/10.15587/1729-4061.2021.239253
19. Amirgaliyev, Y., Ataniyazova, A., Buribayev, Z., Zhassuzak, M., Urmashev, B., Cherikbayeva, L. (2024). Application of neural networks ensemble method for the Kazakh sign language recognition. Bulletin of Electrical Engineering and Informatics, 13 (5), 3275–3287. https://doi.org/10.11591/eei.v13i5.7803
20. Dey, A., Biswas, S., Le, D.-N. (2024). Recognition of Wh-Question Sign Gestures in Video Streams using an Attention Driven C3D-BiLSTM Network. Procedia Computer Science, 235, 2920–2931. https://doi.org/10.1016/j.procs.2024.04.276

21. Satybaldiyeva, R., Uskenbayeva, R., Moldagulova, A., Kalpeyeva, Z., Aitim, A. (2019). Features of Administrative and Management Processes Modeling. Optimization of Complex Systems: Theory, Models, Algorithms and Applications, 842–849. https://doi.org/10.1007/978-3-030-21803-4_84
22. Athira, P. K., Sruthi, C. J., Lijiya, A. (2022). A Signer Independent Sign Language Recognition with Co-articulation Elimination from Live Videos: An Indian Scenario. Journal of King Saud University - Computer and Information Sciences, 34 (3), 771–781. https://doi.org/10.1016/j.jksuci.2019.05.002
23. Rao, G. A., Kishore, P. V. V. (2018). Selfie video based continuous Indian sign language recognition system. Ain Shams Engineering Journal, 9 (4), 1929–1939. https://doi.org/10.1016/j.asej.2016.10.013
24. Aitim, A. K., Satybaldiyeva, R. Zh., Wojcik, W. (2020). The construction of the Kazakh language thesauri in automatic word processing system. Proceedings of the 6th International Conference on Engineering & MIS 2020, 1–4. https://doi.org/10.1145/3410352.3410789
25. Nuralin, M., Daineko, Y., Aljawarneh, S., Tsoy, D., Ipalakova, M. (2024). The real-time hand and object recognition for virtual interaction. PeerJ Computer Science, 10, e2110. https://doi.org/10.7717/peerj-cs.2110
26. Kolesnikova, K., Mezentseva, O., Savielieva, O. (2019). Modeling of Decision Making Strategies In Management of Steelmaking Processes. 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), 455–460. https://doi.org/10.1109/atit49449.2019.9030524