American Sign Language Recognition for Underrepresented Populations

Taylor Balsky
Queen's University
taylor.balsky@queensu.ca

Jeffrey Di Perna
Queen's University
22dngb@queensu.ca

Shreya Menon
Queen's University
22sm47@queensu.ca

Brian Perez
Queen's University
brian.perez@queensu.ca

Shrika Vejandla
Queen's University
shrika.vejandla@queensu.ca

Annie Wu
Queen's University
23dbf1@queensu.ca

Wendy Zhang
Queen's University
zhang.wendy@queensu.ca

Abstract—**Interactive educational platforms for learning standardized material, such as new languages or academic topics, have become increasingly popular. However, American Sign Language (ASL) educational tools remain limited, despite the need for accessible and effective ASL learning resources. Artificial intelligence (AI) advancements in interactive educational applications have greatly improved their functionality and versatility. AI is a highly viable and appropriate approach to creating a tool for ASL learning. In translating between text-based languages, there is a simple and consistent mapping between corresponding words and phrases. ASL requires analysis of spatial and temporal features, making AI integration uniquely challenging. This project explores the limitations of ASL education, particularly in the context of interpreter supports and technology. Our project explores various AI models that can effectively promote ASL learning, and provides experimental results for the implementation of various 2D Convolutional Neural Networks (CNNs). Our research prioritizes ethical considerations by carefully selecting datasets to minimize bias, ensuring that AI-driven ASL tools promote inclusivity and accuracy in sign language learning.**

I. Introduction

Artificial intelligence (AI) poses immense potential in revolutionizing education, particularly with respect to language learning. AI-driven tools have promoted learning spoken and written languages through unique means such as providing real-time feedback on errors and accuracies, acting as personalized instruction. However, the integration of AI into American Sign Language (ASL) education is limited. ASL, as a visually and spatially dynamic language, may potentially require a unique pedagogical approach that traditional educational tools, which are limited as they are, lack, and thereby warrant the integration of AI [Pirone et al., 2023]. The absence of accessible, effective AI-driven ASL education tools limits the ability for learners, including interpreters and individuals who are Deaf or Hard of Hearing (HoH), to promote their learning [Pirone et al., 2023]. The lack of technological innovations in ASL education may impede language acquisition in devaluing the practice of self-reflection for ASL learners, particularly interpreters and educators, diminishing the ability to adapt to individuals' unique signing [Pirone et al., 2023]. Furthermore, the lack of educational tools, particularly those that keep up with advancements at the intersection of technology and education, speaks to the undervaluation of ASL education as an academic discipline, restricting opportunities for communication accessibility and inclusivity in society at large [Pirone et al., 2023].

This project seeks to critically examine narrative thematic findings in the literature surrounding ASL education, investigating opportunities and shortcomings that could potentially be addressed by AI. Furthermore, upon reviewing these limitations, we seek to conduct an exploratory data analysis of the datasets that could be used to develop an equitable AI model that detects and provides feedback on signing. It is crucial to consider the representation of those with disabilities, racialized communities, and those who are natively Deaf and HoH. This project also seeks to appraise models that are able to provide feedback on signing, offering quantitative and qualitative insights into their efficacy. Namely, AI models utilizing 2D Convolutional Neural Networks (CNNs) will be developed to detect and provide feedback on signing. CNNs have demonstrated the ability to facilitate real-time image and video processing, making them well-suited for extracting spatial features from sign language videos.

The implementation of such a model has the potential to promote ASL education by providing real-time corrective feedback, improving interpreter training, and fostering a more inclusive learning environment for all ASL users in the models' representation of those who are racialized, disabled, Deaf, and HoH. Experimental results from numerous CNN implementations will be assessed to determine the most effective model for ASL recognition, with a focus on accuracy optimization so as to minimize bias in recognition performance particularly among diverse signers. The research also harbours ethical considerations associated with AI integration in ASL education, such as dataset biases and the reliability of AI-generated feedback. In addressing these considerations, this project seeks to ensure that AI-based ASL learning tools are accurate and equitable, contributing to more effective ASL education and interpreter training. It is imperative that the benefits of AI extend to ASL learners, helping promote systems that prioritize communication for those with disabilities.

A. Motivation

The focus of this paper is on identifying, analyzing, and addressing the limitations of AI-driven educational tools for American Sign Language (ASL) learning. The narrative thematic analysis of the literature surrounding ASL education examines existing challenges in ASL education, particularly the lack of effective technological solutions, and explores how AI, specifically 2D Convolutional Neural Networks (CNNs), may improve sign recognition and feedback. Few studies have investigated the means by which technological innovations, such as AI-powered ASL learning tools, may be leveraged to promote accessibility, effectiveness, and potential shortcomings in ASL education [Pirone et al., 2023].

This paper is hence motivated to not only examine AI model performance but also consider key aspects such as data collection and diversity in datasets. Sign recognition models, including 2D CNNs, may rely on datasets that may not adequately represent the full range of ASL variations across different signers, leading to biased outputs and reduced accuracy for certain populations, such as racialized populations who may have darker skin. It is hence crucial to ensure that AI-driven ASL education tools are inclusive and that biases can begin to be addressed at the dataset level, ensuring that AI-generated ASL feedback is reliable and holds authentic, pedagogical value without perpetuating inequities in education.

This research contributes to an emerging area in both AI and ASL education, aiming to bridge the gap between technological advancements and practical applications in sign language learning. AI-based ASL must not only demonstrate high accuracy but also impede biases that may diminish accessibility and learning outcomes. This paper seeks to explore a means by which AI-driven ASL education tools can be made technically sound and beneficial for learners, interpreters, and the broader Deaf and HoH community.

B. Problem Definition

Current ASL education tools lack effective AI-driven solutions for real-time feedback, limiting learning for prospective interpreters and individuals who are Deaf or HoH. Traditional language learning platforms may rely on text-based approaches that do not account for the spatial and temporal complexity of ASL. To address this gap, we propose building a machine learning (ML) model that classifies ASL signs from input videos, serving as the foundation for an interactive learning interface. This model is designed to assist users in practicing their signs by providing real-time feedback on accuracy and fluency.

A major challenge in ASL recognition is the ability of AI models to accurately interpret sign language movements while minimizing bias. Many existing ASL datasets lack diversity in signers, which can lead to models that perform inconsistently across different users. Our model will be trained on the Word Level American Sign Language (WLASL) and Microsoft ASL Citizen datasets [Li et al., 2020], [Desai et al., 2023], which may provide a more broad range of signing styles.

To improve sign recognition, this project will explore the use of 2D CNNs to analyze individual frames from ASL video data, extracting spatial and temporal features to classify signs. By identifying, analyzing, and potentially mitigating potential biases in the dataset and model training process, this research aims to enhance the reliability of AI-powered ASL education tools.

II. Background and Related Works

A. Research Questions

1) What are the most effective methods for improving AI-driven ASL recognition and minimizing bias in sign language datasets?
    - One approach involves selecting diverse and representative datasets, such as Word Level American Sign Language (WLASL) and Microsoft ASL Citizen [Li et al., 2020], [Desai et al., 2023], to ensure that the AI model generalizes well across different signers.
    - Model optimization techniques, such as data augmentation and transfer learning, can further improve recognition accuracy.
2) What challenges exist in AI-based ASL education, and how can they be addressed?
    - One key challenge is the accurate interpretation of ASL's spatial and temporal complexity, which requires AI models to process continuous movement rather than static text.
    - Another substantial challenge is the lack of standardized evaluation metrics for AI-driven sign language education tools, making it difficult to assess their effectiveness.
3) How do we integrate Machine Learning (ML) into educational contexts?
    - ML can be applied to educational systems to help users learn new concepts through training exercises. The model's ability to classify can be leveraged as a feedback tool for learners as they practice, providing real-time analysis and suggestions.
4) What are specific methods to enhance model fairness and prediction accuracy?
    - Exploratory Data Analysis (EDA) is a crucial component of the ML pipeline as it provides statistical and visual representation of biases and imbalances in the dataset. Upon performing this step, data augmentation can be done to combat the issues identified.
    - Once a model is created, a confusion matrix can be used to analyze its performance. This gives great visual insight into weaknesses of the model, highlighting its common misclassifications. Modifications can be made to the dataset and model architecture, such as data augmentation, training epochs, and dropout, to enhance overall performance.

B. Contributions

The main contributions of this paper are summarized below:

1) We conduct a narrative thematic analysis approach to identify key limitations in ASL education per the literature, particularly regarding the lack of innovation and standardization of the curriculum.
2) We examine the potential for the *WLASL* and *Microsoft ASL Citizen* datasets to potentially mitigate underrepresentation of diverse populations. We also explore the application of 2D Convolutional Neural Networks (CNNs) for ASL recognition, assessing their potential to improve real-time sign feedback.

C. Related Works

ASL is crucial to communication for those who are Deaf or Hard of Hearing, however, it continues to be established as a rigorous academic discipline, despite gaining immense popularity among the general public in recent years [Pirone et al., 2023].

1) **Shortages of ASL educators and curriculum limitations**
The shortage of qualified ASL educators and the lack of a standardized, research-based curriculum poses significant challenges to ASL education [Pirone et al., 2023]. Unlike spoken languages, ASL is often classified under special education departments instead of being recognized as a typical language department, limiting its scope and ability to be established as its own discipline in mainstream academic programs [Pirone et al., 2023]. Few programs exist for ASL instructors to be properly trained, particularly with respect to terminal degrees, with even fewer allowing specialization in ASL pedagogy, further restricting the quantity of trained instructors [Quinto-Pozos, 2011] [Swaney and Smith, 2017]. Moreover, existing training programs often lack rigorous methodologies, relying primarily on anecdotal claims and unverifiable field testing rather than empirical data [Thoryk, 2010]. Alongside a lack of programs, many ASL educators are not formally trained in language instruction, lack specialized training in ASL education, or hold degrees in related fields rather than in ASL pedagogy or second language acquisition, primarily due to the lack of terminal degrees in ASL education in North America [Pirone et al., 2023].
The absence of a standardized, research-based curriculum further weakens the quality of ASL education, forcing many educators to rely on commercial materials, which often fail to be backed by empirical work [Pirone et al., 2023]. At certain institutions, instructors are restricted to specific curricula, preventing them from tailoring content to teaching methods and student needs [Pirone et al., 2023]. In order to enhance ASL education, [Rosen, 2010] suggest curriculums to include content-based instruction (CBI) and task-based language teaching (TBLT) [Rosen, 2010]. Overreliance on commercial curricula limits instructors' ability to incorporate creativity and self reflection in their teaching [Pirone et al., 2023]. Pirorne et al. emphasize the fact that fluency does not imply one has the ability to teach effectively. Ultimately, though, these gaps in educator preparation and curriculum development speak to the need for innovation and an increased number of opportunities to reliably improve signing as part of ASL education to ultimately serve as an interpreter.
2) **Challenges with ASL assessment and proficiency evaluation [Paludneviciene et al., 2012]**
Alongside a well-developed curriculum, effective ASL education also calls for reliable proficiency assessments. However, researchers argue that establishing clear ASL standards must precede developments in assessments. Presently, sign language proficiency, including ASL, is evaluated through various methods, including the Sign Language Proficiency Interview (SLPI), which evaluates grammar, vocabulary, production, fluency, and comprehension on a 0-5 scale involving 3 raters. Other common approaches include behaviour checklists, performance-based tests on targeted linguistic aspects of ASL, and objective tests (i.e., objective right or wrong evaluations of metrics such as vocabulary and grammar). Despite these assessments, scholars still are unaware how to best evaluate a visual language, which may be adapted by the user differently. One major challenge is that certain tests involve direct translations between English and ASL, despite certain words lacking direct translations. Consequently, test administrators may rely on fingerspelling, which introduces English influences and potentially alter test content. Additionally, many assessment tools have been developed by hearing individuals, prompting scholars to advocate for greater collaboration between sign language linguists, native Deaf signers, and test developers in order to improve the tests' validity and ability to serve Deaf, HoH, and disabled communities. Ultimately, educators require a diverse toolkit of assessments in order to properly evaluate ASL proficiency given the unique structure of the language.
3) **Phonological fluency and expressive skill development in ASL**
Developing fluency in ASL is challenging due to the structure of ASL's morphology that differs fundamentally from spoken languages: namely, morphological structures are encoded simultaneously instead of sequentially [Paludneviciene et al., 2012]. Moreover, ASL involves the use of manual and non-manual articulators, converting multiple layers of meaning simultaneously [Paludneviciene et al., 2012]. Unlike spoken languages, ASL lacks a widely used writing system, making it difficult to capture essential linguistic features such as grammatical inflections, body movements, and effective information [Paludneviciene et al., 2012] [Quinto-Pozos, 2011]. A potential solution is exploring ASL writing systems, as reviewing signed language is significantly

harder than written language, even with the use of recording technologies [Quinto-Pozos, 2011].

While common nouns can be easily translated, complex inflected signs that encode information about direction, number, and subject-object relationships are difficult to represent in written form, which may create barriers for learners [Quinto-Pozos, 2011]. Certain ASL curricula still emphasize individual signs, reinforcing the misconception that ASL consists of signs structured according to English grammar, and that textbooks can be used as vocabulary lists instead of learning material [Quinto-Pozos, 2011].

Researchers suggest curricula to focus on classifiers and constructed action, and how to coordinate the two elements simultaneously [Quinto-Pozos, 2011]. Additionally, instructing iconicity in ASL may be helpful for L2 learners, but research is still needed to confirm effectiveness; similarly, fingerspelling, though often overlooked, may be useful in language development [Quinto-Pozos, 2011].

4) **Systemic barriers in ASL education: Audism, linguicism, and lack of diversity**

Audism, or discrimination against Deaf individuals, has impacted the structural inequalities in ASL education [Pirone et al., 2023]. Such systemic bias can be found in hiring practices, the classroom, and through institutional policy discriminatory against Deaf teachers and students [Pirone et al., 2023]. Deaf teachers have historically been overlooked for faculty positions within the hearing community despite having had more qualifications [Pirone et al., 2023]. Furthermore, Deaf students are also positioned in educational environments built predominantly for hearing students, limiting equal access to resources and professional advancement opportunities [Paludneviciene et al., 2012]. The result is a perpetual cycle whereby Deaf individuals remain underrepresented among the teaching faculty and leadership positions, further validating the notion that hearing teachers are more suitable for academic positions in ASL programs [Swaney and Smith, 2017].

Linguicism describes the discrimination against individuals or groups per their language, and it encompasses the preferential treatment of spoken languages over signed languages, pushing ASL further out of the academic arena. ASL was historically left out of general language courses, with the majority of universities placing it in the category of communication disorders rather than linguistics or foreign languages departments [Rosen, 2010]. This placement de-legitimizes ASL as an academic subject of study and hinders its access to grants and institutional support [Buisson, 2007]. The prejudice has structural issues that limit ASL's educational expansion and accreditation as an autonomous linguistic system.

ASL instructor diversity continues to be a persistent issue, with faculty compositions predominantly white and hearing. Though ASL education has grown exponentially, opportunities for Deaf instructors, particularly minority ones, are still lacking [Pirone et al., 2023]. Institutions typically point to a lack of Deaf professionals holding higher-level qualifications as the reason for underrepresentation among Deaf faculty members, yet there is very little investment in developing chances for Deaf scholars to gain these qualifications [Paludneviciene et al., 2012]. This underrepresentation not only affects employment equity but also the educational environment since students will have fewer opportunities to engage with diverse role models who can provide authentic cultural and linguistic insights. In addition, deaf students and faculty typically experience issues in obtaining necessary accommodations such as interpreters in faculty meetings and research seminars that further place them in exclusion from academic environments [Paludneviciene et al., 2012].

Overcoming such barriers requires active institutional change. Universities must also commit to the hiring of more Deaf educators and professional growth through mentorship programs, graduate school funding, and equitable hiring practices [Swaney and Smith, 2017]. ASL programs must also be situated within language or linguistics departments rather than in special education departments so that ASL is accorded the same respect and resources as other spoken languages [Pirone et al., 2023]. Institutions must also ensure accessibility and provide comprehensive accommodations for Deaf faculty and students to achieve a warm academic environment.

5) **Technology in ASL instruction**

Various technologies have been incorporated into ASL instruction to enhance accessibility and effectiveness without relying solely on traditional in-person instruction [Shao et al., 2020]. Video and digital video disc (DVD) recordings have long been used for ASL education, serving as instructional materials for both learning and assessment [Quinto-Pozos, 2011] [Thoryk, 2010]. Additionally, computer-based programs, such as a DVD program for learning Australian Sign Language (Auslan), highlight the importance of incorporating regional dialect variations into sign language instruction [Ellis et al., 2011]. While these resources provide valuable learning materials, they lack interactivity compared to more advanced technologies that enhance user engagement and experience. One notable advancement is Automatic Sign Language Recognition (ASLR), which has been used to develop tools such as SignQuiz, a quiz-based learning tool for fingerspelling in ISL (Indian Sign Language) [Joy et al., 2020]. Similarly, machine translation technologies have contributed to the development of 3D avatars capable of replicating facial expressions and movements, making sign language learning more accessible and immersive [De Martino et al., 2017] [Papastratis et al., 2021]. [Mehta et al., 2019] further expand on this concept by proposing an automated system for generating 3D sign language video captions, showcasing how AI-driven tools can enhance ASL education.

Recent innovations involve wearable technology, such as smart glasses, which utilize augmented reality and sensor-based capturing to assist Deaf and Hard-of-Hearing students with lecture comprehension [Miller et al., 2017]. Additionally, gesture-capturing technologies, including Kinect and Leap Motion sensors, as well as data gloves, have been used to analyze and facilitate sign language learning [Papastratis et al., 2021]. These tools vary in effectiveness, with some prioritizing accuracy at the cost of computational complexity, while others enable real-time interaction but may lack precision.

Another emerging area is mixed-reality (MR) technology, which enhances ASL learning by incorporating real-time feedback and immersive experiences. Studies have demonstrated the benefits of interactive learning over passive approaches, emphasizing the need for further research into AI-driven ASL systems to integrate advanced feedback mechanisms [Shao et al., 2020].

In the status quo, challenges remain with integrating technology with ASL education. Machine learning-based Sign Language Recognition (SLR) is limited by the scarcity of large, diverse datasets, which affects both recognition accuracy and generalization abilities [Papastratis et al., 2021]. Sign Language Translation (SLT), which involves sequence-based ML algorithms, faces similar dataset limitations that hinder progress [Papastratis et al., 2021]. Despite providing significant potential, continued developments are necessary to overcome these limitations and create more effective, accessible, and interactive ASL learning tools.

III. Methodology

A. EDA & Dataset

In this project, we looked at two datasets, both having upwards of 2,000 classes. The first was *WLASL*, which was composed by Dongxu Li and Hongdong Li for the purpose of benefiting communication between deaf and hearing communities. From an environmental perspective, various backgrounds and lighting conditions are present in the *WLASL* dataset. Regarding the signers themselves, there are over 100 different individuals in the dataset, with each sign performed by at least three of them. Diversity among the signers is also significant, with clear variety in gender, age, and cultural representation. EDA revealed several underrepresented words with much fewer class instances. After creating a graph of the number of videos for each word, or the number of files within each sub-folder (based on the extraction and storage of the dataset), a bar graph displayed that several words had over 14 video examples, whereas the median number of videos per class was approximately four. Underrepresented words lead to the model being biased and less able to recognize those words due to less video data to draw upon and learn from. To remedy this problem, the *MoviePy* library was used to traverse every folder with fewer than four videos and select one of the existing videos to randomly augment (flipping, rotating, changing the brightness, and cropping to a limited extent). We repeated this process until there were at least four videos in each folder (for every word). This process attempts to further enhance diversity into the model, increasing variation in its data to improve its recognition of these words.

The other dataset we looked into was *Microsoft ASL Citizen*, which was developed by Microsoft Research with the help of Boston University, University of Washington, and the Rochester Institute of Technology. It is the first crowdsourced and largest Isolated Sign Language Recognition dataset. Many videos in this dataset are filmed in candid conditions, enhancing the authenticity of the data. This dataset includes more diversity in that it includes signers of various minorities, including those in the Deaf community and those with disabilities. From an ethics perspective, this is a good dataset as explicit consent was received from every contributor.

Though the original goal of the project was to create an educational interface to assist learning signers with a multitude of words and phrases, it became clear that the ratio of videos per class to classes was incredibly small, even with data augmentation. To better comply with the data requirements of this project, the top five most populated classes in the *WLASL* and *Microsoft ASL Citizen* dataset were chosen as the subset to be used in the project. The words were *bite, dark, decide, demand, dog*. These words were assigned labels of *0, 1, 2, 3, 4*, respectively.

B. Experiment Setup

We took on the challenge of training a model with video input. We felt that in signing everyday words and phrases, spacial and temporal features are best conveyed through videos. We initially started researching the I3D (Inflated 3D CNN) model for recognizing the user-performed signs [Haizhong, 2021]. This model is based on analyzing individual images via a 2D CNN architecture and extending it into a 3D CNN by capturing changes between individual frames. The process starts with taking in a video as the user input and dividing it into separate frames. It then examines the images using filters that slide over the height and width to detect objects by identifying changes in color from pixel to pixel, creating an outline for the figure. From there, the model compares the object positions from the previous frame to identify movement within the video. After completing this analysis, the program identifies which sign the user is performing. We sought to leverage a pre-trained model due to limitations in computational power. As well, prior research demonstrated promising results from taking a similar approach [Wong et al., 2022].

Unfortunately, despite our extensive research and high optimism surrounding this pipeline, the dataset size and computational power required to successfully deploy this experiment was shown to be far beyond that available to us. In light of this, we decided to pivot to a new model with greater feasibility given our restrictions.

In shifting gears, we took a new common approach to video classification to execute the task at hand, consisting of a 2D CNN + Recurrent Neural Network (RNN). Through this architecture, the CNN learns spatial features of the video frames (images), and the RNN learns temporal features among the various frames. This process essentially simplifies a 3D problem into two simpler problems in 2D and 1D. A similar roadblock was once again encountered as our available computational power and resources did not allow for proper training on this model. As a result, we were unable to deploy it for our project. With that being said, a simplified version of this architecture was successfully implemented. Removing the LSTM, leaving the model as a 2D CNN proved to be an adequate classifier for this undertaking.

To preprocess this data, we created a dataframe to identify each video and its corresponding enumerated label. Each video was split into frames (30 frames per video). The set of frames for each video was manually analyzed, ensuring only relevant frames were kept. The frames after starting the video but before executing the signs as well as the frames after executing the signs before ending the video were omitted. The new sets of frames were then augmented through cropping, flipping, colour adjustment, and normalization. After that, the dataset was split into training and testing. For each video class, 80% of the videos (i.e., sets of frames) were placed into the training dataset, while the remaining 20% was used as testing data. This way, the train-test split stayed consistent at 80% to 20% on an overall basis and on a per-class basis. From there, the data was trained on four different pre-trained models commonly used for image classification: *ResNet50, InceptionV3, VGG16, MobileNetV3*. Layers were frozen with the exception of the last four so that the models could be fine-tuned on our datasets.

For the graphical user interface (GUI), we envisioned a simple yet efficient sign language recognition system designed to provide users with a clear and interactive experience for real-time interpretation. The interface includes key features such as a functional webcam for live classification, labels displaying predicted signs, and accuracy metrics to offer users feedback on their gestures. Our goal was to create an accessible online solution that is both intuitive and effective. Prioritizing simplicity and accessibility, we identified *Hugging Face* as a viable program. As an open-source platform, it enables seamless deployment of ML models in AI-driven applications.

C. Evaluation Methods

The models were evaluated using the metrics of accuracy, precision, recall, and F1. Confusion matrices were produced for each model to depict the class-by-class breakdown for the predictions. Further data augmentation was performed as input for some of the models if problematic trends were apparent through the confusion matrix.

D. Data Availability

Links to Kaggle are included here.

- *WLASL* Dataset
- *Microsoft ASL Citizen* Dataset

Listed below, through Table I, are the results of the four CNN models used in this project.

TABLE I
RESULTS FROM THE EXPERIMENTS AGAINST THE TEST SET.

Model	Accuracy	Precision	Recall	F1
ResNet50	57.14%	60.20%	57.31%	58.11%
InceptionV3	59.18%	58.83%	58.40%	58.03%
VGG16	81.63%	84.86%	81.96%	81.25%
MobileNetV3	61.22%	64.49%	60.81%	61.31%

Pictured below, through Figure 1 and Figure 2, is the learning curve and confusion matrix of the top-performing CNN model used in this project: the *VGG16* pretrained model.

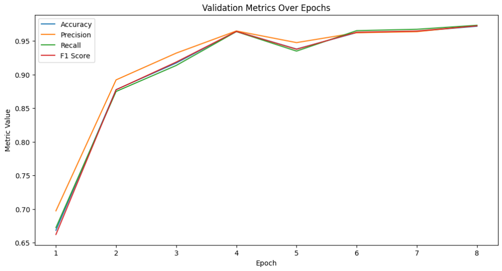

Fig. 1. Learning curve for the *VGG16* model.

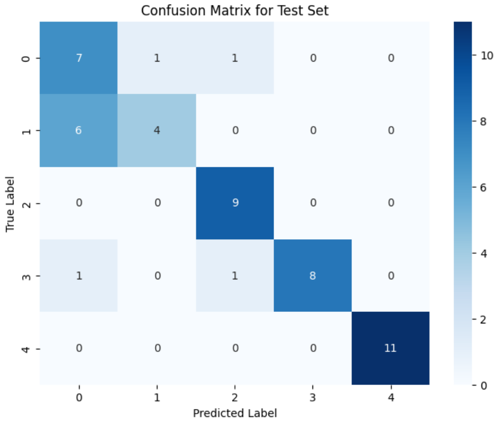

Fig. 2. Confusion matrix for the *VGG16* model.

E. Analysis

This project served as a great educational endeavour as it allowed the team to delve deep into the pipeline of complex computer vision projects and highlighted the significant tradeoffs that can make or break an ML project. Despite the significant challenges faced throughout this process, all stemming from storage and computational power limitations, the team was able to adapt to unideal circumstances and create numerous working and well-performing classification models.

It was clear from the beginning that in taking on a project of this magnitude with the resources available, it would not be feasible to create a decently performing model from scratch. When leveraging the pretrained models, freezing layers and fine-tuning proved to be advantageous. These steps helped the models extract nuanced features of the sign gestures.

When analyzing the achieved results, interesting observations and takeaways are extracted. First, class *4*, *dog*, is shown to have been best classified in all four models, as seen via the confusion matrices. This is due to data augmentation. When we first ran the models, class 4 underperformed relative to the other words. To combat this, we performed further augmentation to enhance the training data for that class. In running the models subsequently, the *dog* class proved to be extremely well-captured through the new, modified dataset.

It is also evident that the *VGG16* model outperformed the rest by a significant margin. This can also be attributed to further data augmentation. When we first ran this model, the trend visible through the confusion matrix was that class *1*, *dark*, was heavily misclassified. As a result, further augmentation was performed on that class. When the model was executed again, it made virtually no errors, aside from further misclassifications of *dark*, but to a reduced extent than before. It showed numerous instances of predicting class *0* on data belonging to class *1*. Though no further refinement to the data was subsequently done, effort in improving the model would entail better distinction between classes *0* and *1* to fix its one consistent mistake.

In analyzing the learning curves for the models in comparison to the metrics obtained when the models was run on the testing data, overfitting is observed. Further modifications to the models would try to address this issue through further optimization of hyperparameters, such as dropout, learning rate, batch size, and epochs.

The models perform slightly below the results achieved in the literature as similar projects obtain results upwards of 80% [Huang and Chouvatut, 2024], [Longlong et al., 2019]. Though we use a small number of classes, we are also limited on the training and testing data we have. In total, we use just under 250 videos, with less than 30 frames per video because manual denoising (i.e., removal of frames from before and after the signing gesture itself) resulted in the discarding of frames. We believe that the biggest limitations our models face is the size of the dataset and the quality of the data and models. The data does not lack quality from a diversity standpoint, but rather in regards to its features. We believe that using a higher framerate would yield better results as there would be more data to train and test on, while the features of the video would be better extracted. As the LSTM is omitted from the architecture, the temporal element of the models is lost. To compensate, a higher framerate would allow for deeper feature extraction in the spacial dimension.

Overall, the models' performance is consistent with the sophistication of the dataset and architecture used, and provides promising insights into the capabilities of 2D CNNs to perform video classification.

F. Ethical Considerations

The ethical considerations surrounding this project surround bias and the effective, reliable use of AI-based ASL learning tools in educational contexts. A key ethical concern is bias with respect to some many words and signs being unrepresented. Thus, in implementing the models explored in this paper with a broad range of classes (words), there may be biases in the model's performance leading to decreased accuracy for words that are represented less. Even with data augmentation methods, biases may be present. The model's ability to generalize may also be skewed toward more frequently represented signs or groups, potentially leading to underperformance in recognizing signs or signs performed by underrepresented demographic groups.

Many of the ethical considerations of this project also speak to the narrative thematic analysis findings summarized in the Related Works section, underscoring the importance of developing a tool that can address the multifaceted limitations surrounding ASL education. Given that our tool is intended to be used in educational settings, particularly for beginners learning ASL, a concern that arises is inequity in accessing the tool. If our tool is inaccessible to those without reliable Internet connection, for instance, this may exacerbate existing disparities in education and communication accessibility for individuals who are HoH, Deaf, and may have intersectional identities such as being from low socioeconomic status, or resource-constrained communities.

Additionally, one substantial concern of our tool is that individuals may potentially overrely on the model for learning ASL, deprioritizing real-world interactions with individuals in applying their learnings. Thus, the broader ASL education system that incorporates AI should take into consideration the extent to which the recognition tool serves as a complement to learning ASL, encouraging collaboration between the learner and the technology, ensuring that the learner retains control over their own learning process.

The potential misuse of this tool, such as using CNNs to detect ASL in healthcare settings, also poses a significant challenge if the model miscommunicates ASL between patients and providers, for example. Human oversight would hence be critical to preventing harm. The key takeaway is that this tool should not be used in isolation in high-stakes decisions, but rather as an assistant to human expertise. Moreover, the tool should primarily be used as an at-home supplement to ASL education, in addition to one's learning in real world contexts with other individuals.

IV. Conclusion

To conclude, the objective of this project was to leverage pretrained ML architectures to create a real-time sign language classification model for common ASL words and phrases. This was to be deployed through an interactive interface as an educational program for users to practice their signs as they start to learn ASL. The *ResNet50, InceptionV3, VGG16, MobileNetV3* models were tested using a combined dataset composed of videos from the *WLASL* dataset as well as the *Microsoft ASL Citizen* dataset. The models were trained and tested on five classes: *bite, dark, decide, demand, dog*. The *VGG16* model outperformed the rest, achieving accuracy of 81.63%, precision of 84.86%, recall of 81.96%, and F1 of 81.25%. The results are promising, showing potential for these models to achieve results in the literature.

Factors contributing negatively to the models' performance include overfitting and insufficient extraction of temporal data. With more time and computational power, proposed amendments include further data augmentation, implementation of the LSTM, and an increased frame sampling rate. Ultimately, this undertaking successfully highlighted the gaps in ASL education systems and proposed a working solution rooted in AI to combat these shortcomings.

V. Future Work

The evolving field of sign language recognition continues to call for opportunities for future advancements. Our findings emphasize the role of model optimization in improving recognition accuracy and real-time performance. To broaden the impact of the project, we seek to refine both the models and GUI for greater accessibility and usability. Enhancing adaptability will allow for more accurate recognition across various lighting conditions, skin tones, hand shapes, and signing speeds. Another addition to the project would be to modify the denoising algorithm. Manual denoising was an adequate solution to preprocess the training and testing videos, but is not doable in real-time implementation of the software. This step must be automated before launching this application for real-world use. From a GUI perspective, future implementations include additional interactive features, such as real-time feedback to offer gesture correction to users. Furthermore, it is crucial to implement Human-Computer Interaction (HCI) elements and interventions as potential means of promoting self-reflection and reducing bias on the part of the user, in addition to examining the means by which the model's datasets may embed biases. Thus, consideration of the ways in which the user interacts with the interface is crucial in understanding the many ways through which bias can be introduced in ML deployment. In consideration of HCI factors such as modifying the time it takes for AI feedback to display, providing gesture accuracy metrics and images, as well as the ways in which the AI is represented (e.g., symbol or individual), the design may work better. Additionally, it is important to evaluate our model's performance by running confusion matrix metrics specifically for the model's ability to predict signs for racialized groups, signers with disabilities, and other underrepresented populations. Ultimately, our main focus remains to create a seamless platform that would assist in making sign language education and interpretation more accessible and user-friendly.

VI. Limitations

In addition to the computational limitations discussed above, there are several limitations to consider regarding the dataset. First, due to limited storage and memory as well as minimal samples per class, we worked with a small subset of the *WLASL* and *Microsoft ASL Citizen* datasets. As a result, the extensive diversity of these datasets was not entirely represented. If this program were to be used by signers of a group not represented in the data, or perhaps in a foreign environment, the models would be limited in performance. Overall, the smaller size and lesser diversity of the subset used decreases the generalizability of our models. While data augmentation can be performed for underrepresented features in the dataset, it is not the ideal solution, and does not override the fact that there are populations, words, and environments underrepresented in this dataset and many others alike.

VII. Acknowledgments

The team would like to extend our sincerest thanks to Sarah Nassar, the director of design for the QMIND Healthcare node for her unwavering technical support throughout this project. The team would also like to extend our sincerest thanks to Paul Wu, the director of design for the QMIND AI Ethics node for his unwavering support throughout this project in helping with the research and ethical components of the deliverable.

REFERENCES

[Buisson, 2007] Buisson, G. J. (2007). Using online glossing lessons for accelerated instruction in asl for preservice deaf education majors.

[De Martino et al., 2017] De Martino, J. M., Silva, I. R., Bolognini, C. Z., Costa, P. D., Kumada, K. M., Coradine, L. C., Brito, P. H., Amaral, W. M., Benetti, A. B., Poeta, E. T., Angare, L. M., Ferreira, C. M., and De Conti, D. F. (2017). Signing avatars: making education more inclusive. volume 16, Berlin, Heidelberg. Springer-Verlag.

[Desai et al., 2023] Desai, A., Berger, L., Minakov, F. O., Milan, V., Singh, C., Pumphrey, K., Ladner, R. E., Daumé III, H., Lu, A. X., Caselli, N., and Bragg, D. (2023). Asl citizen: A community-sourced dataset for advancing isolated sign language recognition.

[Ellis et al., 2011] Ellis, K., Ray, N., and Howard, C. (2011). Learning a physical skill via a computer: a case study exploring australian sign language.

[Haizhong, 2021] Haizhong, Q. (2021). I3dl an improved three-dimensional cnn model on hyperspectral remote sensing image classification.

[Huang and Chouvatut, 2024] Huang, J. and Chouvatut, V. (2024). Video-based sign language recognition via resnet and lstm network.

[Joy et al., 2020] Joy, J., Balakrishnan, K., and Madhavankutty, S. (2020). Developing a bilingual mobile dictionary for indian sign language and gathering users experience with signdict.

[Li et al., 2020] Li, D., Opazo, C. R., Yu, X., and Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. Snowmass, CO, USA. IEEE.

[Longlong et al., 2019] Longlong, J., Vahdani, E., Huenerfauth, M., and Tian, Y. (2019). Recognizing american sign language manual signs from rgb-d videos.

[Mehta et al., 2019] Mehta, N., Pai, S., and Singh, S. (2019). Automated 3d sign language caption generation for video.

[Miller et al., 2017] Miller, A., Miller, A., Malasig, J., Castro, B., Hanson, V. L., Nicolau, H., and Brandão, A. (2017). The use of smart glasses for lecture comprehension by deaf and hard of hearing students.

[Paludneviciene et al., 2012] Paludneviciene, R., Hauser, P. C., Daggett, D. J., and Kurz, K. (2012). Issues and trends in sign language assessment.

[Papastratis et al., 2021] Papastratis, I., Chatzikonstantinou, C., Konstantinidis, D., Dimitropoulos, K., and Daras, P. (2021). Artificial intelligence technologies for sign language.

[Pirone et al., 2023] Pirone, J. S., Pudans-Smith, K. K., Ivy, T., and Listman, J. D. (2023). The landscape of american sign language education.

[Quinto-Pozos, 2011] Quinto-Pozos, D. (2011). Teaching american sign language to hearing adult learners.

[Rosen, 2010] Rosen, R. S. (2010). American sign language curricula: A review.

[Shao et al., 2020] Shao, Q., Sniffen, A., Blanchet, J., Hillis, M. E., Shi, X., Haris, T. K., Liu, J., Lamberton, J., Malzkuhn, M., Quandt, L. C., Mahoney, J., Kraemer, D. J. M., Zhou, X., and Balkcom, D. (2020). Teaching american sign language in mixed reality.

[Swaney and Smith, 2017] Swaney, M. G. and Smith, D. H. (2017). Perceived gaps and the use of supplemental materials in postsecondary american sign language curricula.

[Thoryk, 2010] Thoryk, R. (2010). A call for improvement: The need for research-based materials in american sign language education.

[Wong et al., 2022] Wong, R., Camgöz, N. C., and Bowden, R. (2022). Hierarchical i3d for sign spotting.