Automated Infant Cue Classification

A Machine Learning Approach to Detecting Hunger and Feeding Discomfort Behavioral Cues in Preterm Infants

Joris Staeb

Utrecht University

MSc Artificial Intelligence
Student number: 6269524
j.stab@students.uu.nl joris.staeb@gmail.com

Project Supervisor	Dr. R.W. Poppe
Second Examiner	Prof. Dr. A.A. Salah
Daily Supervisor	M. Doyran

Utrecht, June 29, 2023

Abstract

Preterm infants, commonly admitted to hospitals, require intensive and time consuming monitoring. Automated monitoring techniques using video data exist but are not being applied in practice for monitoring infants. By allowing automated monitoring through behavioral cue classification, which infants use to communicate their needs, the burden of monitoring can be relieved allowing for improved health outcomes in preterm infants. This study aims to apply machine learning techniques for automated behavioral cue classification in preterm infants to infer their care needs, specifically of hunger and feeding discomfort. A MoViNet model was trained for this classification problem, selected for their extensive pretraining and multi-class video classification capabilities. Due to the limited availability of labeled data, the techniques of few-shot learning and active learning have been applied to investigate if they improve upon baseline performance. Few-shot learning consists of an initial training phase on similar tasks to allow for quick adjustment of weights. Active learning incorporates additional data labeling, with instances gathered using stratified sampling included in the dataset. It was found that the fully supervised baseline approach was able to successfully uncover patterns in infant behavior. However, few-shot learning resulted in worse performance due to challenges in generalizing from the source to the target domain. Active learning performed comparably to the baseline approach and offered additional value as a labeling tool in the data-scarce setting. The research also revealed the impact of individual differences in behavior, affecting the generalizability of behaviors to other infants and hindering performance. Despite these challenges, individual behavioral differences did not entirely prevent successful classification. By incorporating more training data from new infants, the generalizability of the results and performance could be improved. In sum, this research forms a solid foundation for advancing fully automatic infant monitoring, potentially enabling more individualized care with beneficial health outcomes.

Keywords— Preterm infants, Hunger, Feeding discomfort, Behavioral cue classification, Computer vision, Few-shot learning, Active learning, MoViNet

Acknowledgement

I would like to express my gratitude to my project supervisor, Dr. Ronald Poppe, whose invaluable guidance, expertise, and feedback greatly contributed to the realization of this research. I am equally grateful to my second examiner, Prof. Dr. Albert Salah, and daily supervisor Metehan Doyran for their insightful feedback and time. Special thanks to Dr. Jeroen Dudink and the department of neonatology at UMC Utrecht. Their assistance, guidance and willingness to collaborate were crucial to gathering the data for this study.

I am also thankful for the significant contributions of the project team members, Morris Verholt, Eline de Groot, Demi Ennen and Eline Chin. Special thanks go to them for their involvement in the video recording process, as well as helping me navigate the field of neonatal care. I appreciate their time, effort and expertise, which were pivotal to the successful execution of this study.

This thesis would not have been possible without your collective efforts and support, thank you.

Contents

Abbreviations

AI Artificial Intelligence

AL Active Learning

CNN Convolutional Neural Network

FSL Few-Shot Learning

GA Gestational Age

NICU Neonatal Intensive Care Unit

SLAPI Sleep Assessment in Preterm Infants

1 Introduction

1.1 Problem Statement

Premature birth remains a worldwide health concern, with approximately 11% of all births being premature [Vogel et al., 2018]. Despite the neonatal mortality rates steadily declining, the incidence of preterm birth has not followed this trend [Cao et al., 2022]. Consequently, premature birth is an increasingly large contributor to neonatal mortality. Other than increased mortality risk, short-term health challenges are increased risk of respiratory disease, necrotizing enterocolitis and sepsis [Lumley, 2003]. Beyond immediate health risks, premature birth has also been associated with neurodevelopmental problems, more frequent hospitalizations and social-developmental struggles [Vogel et al., 2018].

Premature infants require specialized care, and are therefore commonly admitted to the Neonatal Intensive Care Unit (NICU). Other than monitoring their general health, infants are also taken care of in their primary needs, like sleeping and feeding. The management and timing of feeding affect further development. Evidence suggests that appropriate nutrition stimulates healthy growth development [Su, 2014] and the development of normal oral feeding behavior [Kirk et al., 2007]. Feeding practices commonly fall into two categories: scheduled or cue-based feeding. Scheduled feeding is based on a fixed schedule provided by the practitioner or caregivers. This is the current standard practice, because continuously monitoring all infants for hunger is not feasible with the current demands already posed on nurses [Chrupcala et al., 2015]. However, this approach was shown to cause more stress in infants, as feeding is not always timely and can therefore cause feeding discomfort [Kurt Sezer and Küçükoğlu, 2020]. Cue-based feeding, on the other hand, is guided by behavioral or physiological cues by the infant. Research has shown that cue-based feeding leads to better digestion of food and earlier discharge from the NICU, which has beneficial health outcomes [Wellington and Perlman, 2015; Zimmerman, 2013]. The challenge lies with balancing the demands on the nursing staff and the feeding needs of the infants. Thus, strategies aimed at enabling cue-based feeding should be explored.

By specifically examining the cues of hunger and feeding discomfort, it is possible to identify and address some underlying causes that are detrimental to the health and development of preterm infants. Traditionally, infant hunger has been monitored through physiological measurements or by healthcare professionals through behavioral assessments of feeding behavior and clinical signs such as crying, sucking, and swallowing. However, physiological measurements require invasive tools that can impede with care. Behavioral assessments are prone to subjectivity, and do not provide continuous monitoring of infant hunger and feeding discomfort. It would be beneficial to keep track of the infants' states through automated measures. Research has shown computer vision techniques can be implemented to keep track of infants [Olsen et al., 2014; Huang et al., 2022; Li et al., 2021; Nagy et al., 2021], and this may be exploited to the benefit of an infant by monitoring when it signals it may be in a hunger state and requires feeding. Another key advantage of computer vision techniques is they are non-invasive, which may reduce stress on the already heavily tubed and touched infant.

Machine learning is very suited to visual analysis tasks and has been widely applied [Guo et al., 2016]. Example applications are human pose estimation, tracking, object detection and classification, and temporal action localization. In the medical field, machine learning has been applied for the development of targeted treatments by analyzing large amounts of patient data and identifying unique patterns and trends. This information can be used to create personalized therapies or care routines that are specifically designed for the individual needs of each patient [Haleem et al., 2019]. This is very similar to problem at hand in this thesis, as such the Artificial Intelligence (AI) domain is suited for infant monitoring.

In sum, to best improve the infant care on the NICU, a system should be implemented for the continuous monitoring of hunger and feeding discomfort states in infants. This system should be robust for individual differences in infants and their behaviors

1.2 Scope

This research aims to address the challenges in infant care by examining computer vision techniques suited to detecting hunger and feeding discomfort cues. By automating monitoring of these processes, this research aims to improve the overall well-being and development of infants on the NICU. To achieve this goal, an initial review of the existing literature on infant hunger and automated monitoring systems to identify current knowledge on the topic will be conducted. Then a system must be implemented to test and experiment with the behavioral cue classification. Finally, a discussion of the feasibility and benefit of the implementation must be held.

Due to the use of computer vision techniques to detect behavioral cues, video cameras will be used to record preterm infants in NICU beds. As preterm infants typically start showing behavioral cues for feeding after 32 week Gestational Age (GA) [Whetten, 2016], the recorded population starts at 32 weeks GA. The recordings were made under normal circumstances, and infants that were recorded, were not treated differently. Recordings for this thesis were made available by the Sleep Assessment in Preterm Infants (SLAPI) research project at UMC Utrecht. Only limited data exists, as videos are still being recorded at the time of writing by UMC Utrecht, and additional datasets are not available.

The scarcity of the data limits the range of choices for computer vision architectures that can be successfully trained. Two architectures were specifically designed for handling scarce, and unlabeled data. A Few-Shot Learning (FSL) model is trained on only a small number of examples for an outcome class in a different domain, and is expected to generalize to new examples with only limited additional training data. Active Learning (AL) iteratively trains on the available annotated data, and then selects samples from a pool of unlabeled data for an oracle to label. These architectures will be compared with a standard fully supervised approach. The use of these two techniques is justified by the fact they both approach the issue of the data scarcity from a different angle. FSL maximizes the use of the available labeled instances, possibly using transfer learning. AL learns most from the unlabeled data pool, using the annotated data as reference point. In Section 2.3, these methods and their application in related works are discussed in more detail.

To summarize, this research aims to implement FSL and AL on preterm infant videos for detecting hunger and discomfort cues. It must also be taken into account how differences across infants are challenging to the implementation of a uniform system. The main contributions of this thesis are threefold: (1) the use of machine learning techniques for behavioral cue classification, (2) a the systematic investigation into these cues and individual differences in infants, and finally (3) the compilation of a labeled dataset.

1.3 Research Questions

Given the scope of this thesis framing, the following main research question arises:

> *Can we detect and classify infant hunger and feeding discomfort cues in preterm infants using machine learning?*

By implementing and testing different machine learning approaches, this research aims to diversify literature on infant cue detection. It is hypothesized that promising outcomes are possible in cue classification. This is supported by the work of Sun et al. [2019] and Sun et al. [2021], who classified infant discomfort states based on video data. Since different approaches will be tested, this means the question cannot be answered directly. For each machine learning approach, the same question is asked, with the aim of finding out which approach works best. The sub-questions are shown below:

> *SQ1: Can we detect and classify infant hunger and feeding discomfort cues in preterm infants with a model trained using few-shot learning?*

> *SQ2: Can we detect and classify infant hunger and feeding discomfort cues in preterm infants with a model trained using active learning?*

For *SQ1* it is hypothesized that a meta-learning phase in the FSL pipeline could successfully pretrain the model to leave a more optimized parameter search space for the meta-testing phase, allowing to adequately distinguish the cues based on intricate details in the movement of the infant. This was previously done to distinguish between human motion [Gui et al., 2018] and infant or adults faces [Atallah et al., 2022]. For *SQ2* it is hypothesized that AL would solve the issue of limited labeled data by seeking out the most informative instances and enhance performance, as was previously done by Yang et al. [2015] in a similar multi-class classification problem.

Additionally, individual differences in behavioral cues, including their onset, intensity, and frequency, may exist between infants. It is worthwhile to investigate further how differences between infants affect automatic detection using machine learning. This results in the final sub-question, listed below:

> *SQ3: Do individual differences in infants and cues affect automated detection and classification of behavioral cues?*

To answer this question, it must be determined how model performance is affected by the variation in the data. It is hypothesized that individual differences do affect classification. Research showed that behaviour in infants differs in onset, intensity and form [Thoman and Whitney, 1990; Frischen et al., 2007; Claessens et al., 2011], and that differences in feeding cues affect the success of feeding [Ventura and Mennella, 2017].

Performance will be assessed by using standard metrics: accuracy, precision, recall and F_1. Adequate detection and classification is then a function of how well the output of the methods overlap with the assessment of the medical professionals currently responsible for infants on the NICU and to what extent their responsibilities could be relieved by this system.

1.4 Outline

The structure of this thesis is as follows: Section 2 reviews background literature and related work focusing on preterm infants, behavioral cues, and relevant AI techniques. Subsequently, Section 3 provides a detailed overview of the methodologies applied in this study. The experimental design, explained in Section 4, sets the stage for the presentation of results in Section 5. Section 6 provides a discussion and interpretation of these results, highlighting limitations and potential directions for future research. The thesis is concluded by summarizing the key contributions and offering a forward-looking perspective in Section 7.

2 Literature

In this section the relevant background literature and related works are discussed. First, some background on preterm born infants is provided, focusing on the relevant aspects of their stay on the NICU. This includes hunger and feeding discomfort. Next, automated cue classification and challenges are discussed, including strategies for cue localization in untrimmed videos. Finally, a machine learning approach to this cue classification problem is discussed.

2.1 Preterm Infant

An birth is considered premature if the infant is born before 37 weeks GA [Platt, 2014]. GA measures the length of the pregnancy, starting from the last menstruation of the mother to the birth of the child. Premature birth can be categorized as extreme (<28 weeks GA), very preterm (28 to 32 weeks GA) and moderate (32 to 37 weeks GA).

Premature birth is associated with various adverse health conditions. Mortality rates in preterm infants are higher than in term born infants [Cao et al., 2022]. Surviving preterm infants face a higher risk for cognitive, social and emotional developmental defects [Lumley, 2003]. For example, these infants have a higher incidence of respiratory and intestinal disorders, and the likelihood of a favorable outcome declines with a lower GA [Platt, 2014]. Children that were born preterm are more likely to require special educational attention, with up to 7% suffering from severe cognitive impairment, with the risk of an IQ below 70. To address developmental disorders due to poor support in the vulnerable early stages of life, preterm infants are commonly admitted to the NICU.

One contributor to the developmental problems that preterm infants face is sleep deprivation, given the important role of sleep in their development ex utero. Term born infants sleep on average 22 hours per day in utero, while preterm infants on the NICU sleep only about 15 hours per day [Orsi et al., 2015]. Since sleep is linked to neurological development and development of the central nervous system, this is a troublesome difference [Bertelle et al., 2007]. Disturbances such as bedside care, noise from the NICU and other stimuli may interrupt the sleep cycle [Park, 2020].

Another cause of these developmental issues is sub-optimal nutrition. Conditions such as early onset puberty, lower average height and blood pressure are some issues directly linked to poor nutrition [Embleton, 2013]. It is important to prevent hunger by monitoring cues signaling their need for feeding, as well as assessing infant comfort after feeding. If feeding causes discomfort, that may be an indication that the infant was not properly supported, not fed the right volume or the food was not prepared correctly. Such feeding discomfort may disturb the infant's already precarious sleep cycles, cause exhaustion, result in weight issues and teach poor feeding habits [Fanaro, 2013]. While strategies to prevent feeding discomfort exist, few are medically validated, highlighting the potential use of computer vision-based improvements. Collectively, these issues currently pose problems on the NICU and contribute to sub-optimal feeding practices [Hung et al., 2013].

Infant monitoring on the NICU is typically managed by a team of medical professionals, including neonatologists, pediatricians and nurses. Their responsibilities encompass all the care that preterm infants need, ranging from feeding and changing diapers to medical procedures and monitoring vital signs. As discussed, feeding practices have a large impact on the infant's development. Despite their extensive effort, experience and knowledge, feeding on most NICUs is limited to scheduled feeding, which means that an infant is fed on a fixed schedule, regardless of the actual hunger experienced [Newland et al., 2013]. Infants may be fed by tube or orally (formula or breast milk). Su [2014] notes that this approach can lead to excessive energy intake, leading to further averse developmental effects like obesity or feeding discomfort. Cue-based feeding is feeding based on hunger and satiety cues given by the infant. Studies found that cue-based feeding improves feeding outcomes, and leads to quicker discharge from the NICU [Puckett et al., 2008; Kirk et al., 2007]. So, communication of the infants' needs happens through behavioral cues, and a cue-based feeding approach leads to improved health and quicker discharge from the NICU. Therefore, behavioral cues of hunger and feeding discomfort are of great importance to further advance infant care.

2.1.1 Hunger

Hunger in preterm infants refers to a state of need for nutrition. Since preterm infants are typically less well-developed compared to full term infants, they may have trouble self-regulating their hunger needs and cues [Nyqvist, 2008]. Despite this, their nutrition is vital for their growth and development. When using cue-based feeding, infants are fed more appropriately and are often discharged from NICU earlier [Wellington and Perlman, 2015; Zimmerman, 2013]. Hunger cues from infants indicate their readiness for feeding. Beyond behavioral cues, hunger may be signified by physiological signs such as changes in blood sugar levels, heart rate, and oxygen saturation [McFadden et al., 2021]. Deviations from standard values can then indicate hunger.

Since behavioral cues from preterm infants are indicators of their hunger state and it is relevant to examine how cues can guide feeding practices. After roughly 32 weeks GA, infants show behavioral cues related to feeding [Whetten, 2016]. In literature researching cue-based feeding, the conclusion is most often that the cue-based feeding approach leads to the best developmental results, therefore validating the importance of these cues [Kamran et al., 2020; Zimmerman, 2013; Settle and Francis, 2019]. A ground truth for hunger in infants is currently determined by their response to feeding, and their vital signs. The volume of food they are provided may then be a measure of hunger, and this information is meticulously logged in the SLAPI dataset. However, this also depends on the GA and development of an infant, as larger infants require more food [Liotto et al., 2020]. If an infant shows disengagement or discomfort to being fed, then it may not be hungry after all. Since hunger is a subjective state, it is difficult to determine a golden standard.

NICU nurses working at the UMC Utrecht indicated that hunger cues largely center around the mouth. Any mouth or tongue movements, as well as sucking behaviors are associated with hunger. However, the occurrence of a cue does not always explicitly indicate hunger. Cues may originate from multiple states, like hunger, pain or sleepiness. Some cues may signal urgent hunger, while others indicate only the starting phases of hunger. In the literature, some attempt has been made to categorize these cues as early, active or late. Early cues mean that the infant is transitioning to the state of hunger, and do not require immediate feeding. Active cues indicate that feeding may be provided. Late cues indicate that the ideal feeding moment has passed, and the infant needs feeding as soon as possible. Watson and McGuire [2016] indicate that crying is a late cue. Hodges et al. [2013] state overt cues with obvious negative affect such as crying and being unsettled

Cue	Type	Urgency	Duration
Quiet wakefulness	State	Active	Long
Alertness	State	Active	Long
Crying	State	Late	Long
Mouthing	Motor face	Early	Long
Tongue poking	Motor face	Early	Short
Taut tongue	Motor face	Early	Short
Smacking lips	Motor face	Early	Short
Gazing	Motor face	Early	Long
Head movements	Motor face	Early	Long
Rooting	Motor face	Active	Long
Stirring	Motor body	Active	Long
Arm waving	Motor body	Active	Short
Flexing	Motor body	Active	Short
Hand-to-mouth	Motor body	Active	Short
Reaching	Motor body	Active	Short
Fussing	Motor body	Late	Long
Unsettled	Motor body	Late	Long
Kicking	Motor body	Late	Short
Stable vitals	Physiological	Early	Long

Table 2.1: Hunger cues preterm infants may display to indicate that feeding is appropriate according to systematic reviews [Puckett et al., 2008; McFadden et al., 2021; Watson and McGuire, 2016; Talej et al., 2022; Fry et al., 2018]. Urgency based on research by Hodges et al. [2013] is only tentative as it is for older infants. Duration is only an indication based on how long behaviors approximately take. Short behavior takes seconds while long behavior can last up to minutes.

or irritable are late cues. Finally, subtle cues and cues that are oral in nature (called "*Motor mouth*") are often early cues [Hodges et al., 2013, 2016]. This categorization is only tentative, as reviews and studies often do not report this.

Reviews on hunger have identified consist cues in infants [Puckett et al., 2008; McFadden et al., 2021; Watson and McGuire, 2016; Talej et al., 2022; Fry et al., 2018]. They summarise cues as shown in Table 2.1. Generally, hunger cues can be divided into four categories: state cues, motor cues in the face, motor cues on the body and physiological cues. State cues indicate whether an infant is alert, sleepy, drowsy or crying. Transitional states, such as transitioning from alert to sleepy, are also considered cues. Cues about infant states have been used successfully to determine whether they may want to be fed [Griffith et al., 2017]. For example, states of "*Quiet wakefulness*" and "*Alertness*" signal hunger, and transitions from states of rest like "*Quiet sleep*" to fuzzy states like "*Crying*" do as well. Motor cues are physical movement, and they are categorized into facial and bodily cues. Facial cues are restricted only to movements in the face, since they are very diverse. They include eye gazes, but also chin, mouth, tongue and brow movements. Body cues include full body movements, or any movement in the limbs. This include the tone of the infant, which is the level of tension in the muscles. An infant has low tone if muscles are limp, and high tone when it is more rigid. This also includes limb flexing [Blauer and Gerstmann, 1998]. Physiological cues include the vital functions of the infant, like the previously mentioned blood sugar levels, heart rate and oxygen saturation.

As for automatic detection of these cues, some are not suited given the computer vision approach. Physiological cues oftentimes cannot be recorded using cameras. Cues must also be somewhat general, as individual differences between infants should not affect performance such that a model cannot learn underlying features. Furthermore, some cues might not exclusively signal hunger but also other states. For example, the state "*Crying*" can signal hunger, feeding discomfort and pain. This makes it difficult to determine to which one it belongs, and it is best that cues exclusively signal a hunger state. It is also potentially informative to judge cues by their urgency. However, previous research on hunger cues has not shown definitively how often cues should occur for an infant to be hungry or in what patterns and combinations they do occur.

2.1.2 Feeding Discomfort and Pain

Feeding discomfort is the physical or emotional discomfort that infants can experience during or after feeding [Newland et al., 2013; Shaker, 2013]. It can range from physical to emotional distress, and it can occur during or after feeding. The discomfort can be caused by factors ranging from wrong positioning or support during feeding, overfeeding, digestive issues or discomfort with the type of food being fed. The ability to accurately assess feeding discomfort can contribute to improved feeding practices by adapting feeding approaches in terms of volume or method of feeding.

Pain is different from feeding discomfort as it is a more general physical or emotional state [Fuller, 1996]. Pain may be caused by disease, medical procedures or injuries, whereas feeding discomfort is only caused by factors related to the feeding. However, the more extensive research into pain can offer understanding in behavioral cues that are applicable to feeding discomfort. In their review, Zamzmi et al. [2019] describe how pain expressions can be automatically analyzed using machine learning approaches, using support vector machine and Convolutional Neural Network (CNN) classifiers. In their later study on neonatal pain assessment, Zamzmi et al. [2022] developed a system for detecting behavioral indications of pain. They successfully coded cues from the NFCS, a scientific scale for analyzing facial expressions in newborns, such as "*Eye squeeze*" and "*Bulging brow*" based on rules derived from landmarks detected with ZFace (see Section 2.2.1). Given these more advanced use cases of computer vision techniques, there is potential to apply these approaches to the detection of feeding discomfort and possibly hunger cues. Especially since methodologically the setups are suited to measuring behavioral cues, like the cues seen in Table 2.3. The table include cues from pain scales like the NFCS, but also PAIN and NIPS [Pereira et al., 1999; Grunau et al., 1998; Hudson-Barr et al., 2002] Additionally, cues from different reviews and research papers were also included [Holsti et al., 2005; Morison et al., 2003; Hatfield and Ely, 2015]. To ensure consistency, these cues are categorized into state cues, facial motor cues, body motor cues and physiological cues.

Feeding discomfort can manifest through behavioral cues following the same categorization of hunger cues in state cues, motor cues of the face and body and physiological cues. Thoyre et al. [2013] gives a descriptive account of a feeding case study, where the challenges during and after feeding are taken into account. They observed that after overfeeding the physiological state of the infant became unstable, including other behaviors that can be associated with discomfort like side-to-side movements, coughing, choking and a change in tone. Since these cues directly follow the feeding, it can be concluded they were caused by the feeding and signal feeding discomfort. Similarly, Hung et al. [2013] researched how infants responded to thawed and fresh breast milk. The research showed that after feeding thawed breast milk, preterm infants showed more stress and discomfort cues like retching, choking, change in sleep state, continuous burping, crying, chin quiver and body twisting. Newland et al. [2013] showed that after disengagement cues were not properly picked up during feeding, infants showed discomfort cues of crying, brow lifting, change in tone, splaying of fingers. So these cues clearly indicate that feeding must be improved during the next feeding moment for better digestion and more stable states.

The studies from Newland et al. [2013], Hung et al. [2013], and Thoyre et al. [2013] each highlight instances where disruptions or issues in the feeding process led to discomfort in infants, confirming the importance of the feeding process. These findings were corroborated by NICU nurses at the UMC Utrecht, who recognised feeding discomfort by restlessness all over the body, and that creating a fist, frowning, arching and being unsettled are signs of discomfort. The behavioral cues that were derived from these studies are in shown in Table 2.2.

Cue	**Type**	**Urgency**	**Duration**
Change in alertness	State	Low	Long
Change to drowsy	State	Low	Long
Fatigued	State	Low	Long
Crying	State	High	Long
Uncoupling face	Motor face	High	Long
Incomplete swallow	Motor face	Low	Short
Coughing	Motor face	High	Short
Continuous burping	Motor face	High	Long
Choking	Motor face	High	Short
Brow lifting	Motor face	Low	Short
Change in movement patterns	Motor body	Low	Long
Finger splay	Motor body	Low	Short
Side-to-side movements	Motor body	Low	Long
Body twisting	Motor body	High	Long
Chin quiver	Motor body	Low	Short
Change in tone	Motor body	Low	Long
Increased breathing	Physiological	High	Long
Uncoupled breathing	Physiological	High	Long
Increased heart rate	Physiological	Low	Long
Oxygen desaturation	Physiological	High	Long
Bradycardia	Physiological	High	Long
Apneic	Physiological	High	Long

Table 2.2: Feeding discomfort cues in infants, indicating that food did not settle well and is causing the infant to be uncomfortable [Shaker, 2013, 2017; Thoyre et al., 2005; Hung et al., 2013; Newland et al., 2013]. Urgency only an indication based on how painful behavior must be to elicit cue. Duration only an indication based on how long behavior approximately take. Short behavior takes second while long behavior takes up to minutes.

Cue	Type	Urgency	Duration
Crying	State	High	Long
Irritability	State	Low	Long
Mouth stretch	Motor face	Low	Short
Open lips	Motor face	Low	Short
Upper lip raised	Motor face	Low	Short
Tongue protrusion	Motor face	Low	Short
Tongue extension	Motor face	Low	Short
Taut tongue	Motor face	Low	Short
Coughs	Motor face	High	Short
Chokes	Motor face	High	Short
Gaze aversion	Motor face	Low	Long
Eye squeeze	Motor face	High	Long
Frantic movements	Motor body	High	Long
Extension arms and legs	Motor body	Low	Long
Finger splay	Motor body	Low	Short
Twitches	Motor body	Low	Short
Hand on face	Motor body	Low	Short
Hand-to-mouth	Motor body	Low	Short
Chin quiver	Motor body	Low	Short
Saluting	Motor body	Low	Long
Finger and toe flexion	Motor body	Low	Long
Increased heart rate	Physiological	Low	Long
Oxygen desaturation	Physiological	High	Long
High blood pressure	Physiological	High	Long

Table 2.3: Pain cues in preterm infants. They include cues from pain scales like the NFCS, but also PAIN and NIPS [Pereira et al., 1999; Grunau et al., 1998; Hudson-Barr et al., 2002] Additionally, cues from different reviews and research papers were also included [Holsti et al., 2005; Morison et al., 2003; Hatfield and Ely, 2015].Urgency only an indication based on how painful behavior must be to elicit cue. Duration only an indication based on how long behavior approximately take. Short behavior takes second while long behavior takes up to minutes.

2.2 Infants and Cues in Machine Learning

After providing an overview of the relevant hunger and feeding discomfort cues, it will be examined how previous research has used AI to analyse cues or other behaviors by infants. This includes a discussion of related works on infant tracking, and the measurement of infant behavior. Then, some methods for action detection in videos is discussed, with the aim of providing insight in how infant cues may be detected from live feeds or videos.

Before discussing the literature, the restrictions by the target domain must be noted. Medical datasets are often small, with limited availability. This is challenging to the application of AI techniques such as machine learning, which consume large volumes of data. When the data requirements of such algorithms are not met, it results in a poorly trained models with no application value. Furthermore, the impact of individual differences in cues on classification remains uncertain. An infant could display many early cues, while another may not. An early cue in one infant may be a late cue in another.

The lack of medically validated research in individual differences is a challenge for the interpretation behind the classification.

2.2.1 Infant Tracking

Infant tracking or motion analysis involves studying the movements and behaviors with the purpose of studying development or identifying issues or needs. Within the scope of this research, infant tracking can be achieved through computer vision. Hesse et al. [2019] describe a pipeline as shown in Figure 2.1 for general classification. After acquiring the video data of infants, motion is captured simply through video or additional techniques. Then, motion features can be extracted which are used as inputs for a classifier.

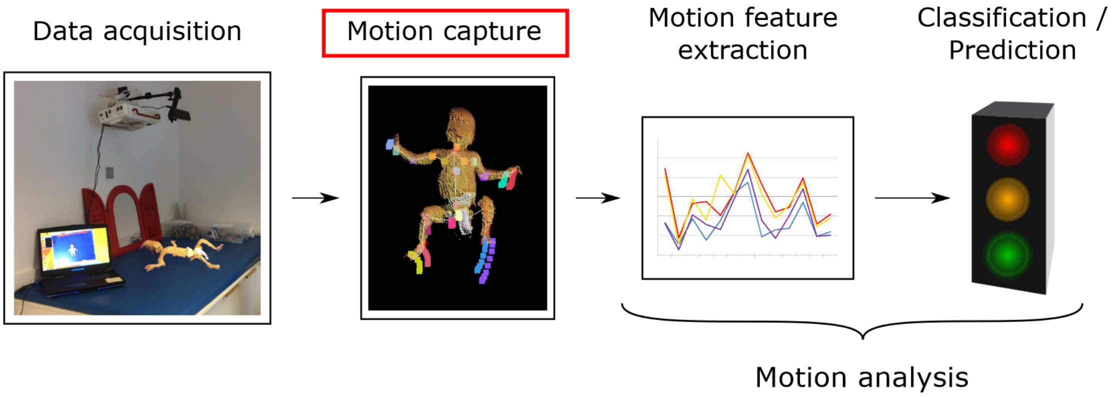


Figure 2.1: Generic motion analysis pipeline. Initially videos are recorded of infants where their motion is captured. Then features are extracted to use for classification or prediction [Hesse et al., 2019].

Infant behaviors and needs can be deduced from cues, therefore implementing automated monitoring relies on measuring these cues from video data. Measurement differs from classification. Cue measurement concerns their presence and quantification, which can then be used for further analysis, such as classification. Cues are categorized into four types: state, motor face, motor body, and physiological. Therefore, the review of related works mirrors this categorization, dividing the tracking process into body and facial tracking.

In the literature, body measurements are often based on landmarking tools designed to mark specific points of interest on the body. Examples of such tools are OpenPose [Cao et al., 2017], EfficientPose [Bukschat and Vetter, 2020] and HigherHRNet [Cheng et al., 2020]. These tools generate a set of two-dimensional points on the image corresponding to detected points of interest. These points can then be processed to determine their movements on a frame-to-frame basis. Points of interest may not be detected if the body is occluded or not adequately shown on video due to poor positioning of the infant. Hesse et al. [2018] and Hesse et al. [2017] successfully measured infant bodies using 3D pose estimations using low-quality RGB and depth data. By capturing the shape of the infant's body, features can more easily be extracted. Olsen et al. [2014] built a 3D model of the infant using depth images to locate the extremities of an infant's anatomy. In practice in the medical field, when attempting to find deficiencies in infants, implementations rely on these full body measurement in order to determine their presence or absence. Chambers et al. [2020] showed that developmental disorders in infants can be assessed using OpenPose augmented for their own training set, after annotating infant videos from the clinical setting. Hashemi et al. [2012] assessed risk of autism spectrum disorder through behavioral markers and their automatic detection for interpreting them on a scale made for infants. From 2D body poses they measured arm asymmetry and other scale specific markers like smoothness of visual tracking and attention disengagement on other objects. Their tools significantly reduce the burden on human assessors as their non-invasive methods match human scores to a clinically satisfactory degree. One recent study did not use such landmarking features for their classifier of lung disease. Navaneeth et al. [2020] used thermal imaging fed to a CNN. They left feature extraction to the network, and as a result detected respiratory syndromes in infants with recall and precision of .92. A final study performed general motion analysis in infants using representation learning [Gong et al., 2022].

Similar to body measurements, facial measurements are often based on facial landmarking techniques. Some packages suited for infant facial landmarking are OpenFace [Baltrusaitis et al., 2018], ZFace [Ertugrul et al., 2019] and InsightFace [Guo and Deng, 2019]. These tools again output landmark coordinates as points on the 2D image, and occlusions prevent accurate measurement. Similar to the aim of this thesis, Sun et al. [2019] did video-based discomfort detection for infants by first doing facial landmark detection and then, after feature transformation, feeding those features to a support vector machine classifier. Their study successfully managed to distinguish between discomfort and comfort states in infants based only the facial information. Ritu et al. [2022] summarizes that facial landmarks are a computationally cheap and non-invasive way to monitor infant states, and can help reduce discomfort in neonates by detecting the source. However, approaches without facial landmarking have also been successful in detecting facial expressions in infants. Li et al. [2021] studied how standard classifiers in combination with image descriptors can be used to measure facial expressions like discomfort, neutral, sleep, joy, open mouth, unhappy, pacifier. They trained on images in these classes, and varied between image descriptors like histogram-oriented gradients and local binary patterns. They also varied their classifiers between CNNs and hidden Markov models, where the latter proved most successful in reducing the number of false positive predictions.

Since most body measurements of infants include landmarking, an issue may occur when landmarks are not detected, and therefore vital information may be missing for classification. An opportunity lies beyond these methods, by possibly using CNNs to detect the relevant features, as the recent papers by Li et al. [2021] and Navaneeth et al. [2020] indicate.

2.2.2 Cue Localization

In order to detect cues and possibly generate more training data, video snippets must be generated from untrimmed video data. In their survey on action detection in untrimmed videos, Vahdani and Tian [2021] investigated the current body of research on action localization. They recognise the difficulty in inputting an untrimmed video directly into a visual encoder due to its size, and this is often solved by using a snippet based approach. In this approach, the untrimmed video is divided into equal partitions. A classifier is subsequently trained using this data. In the classification phase, untrimmed data is provided the the model and is divided into snippets to fit the training phase. However, this partitioning can be performed using fixed boundaries, or more informed boundaries using the trained model to predict where actions start and end.

However, only a small pool of annotated data exists for preterm infant cues. It is potentially valuable to expand on the segments that are annotated. Considering the frames preceding and following the known cues could provide value, as they're likely to contain similar hunger cues. This approach also helps account for cue diversification over time and edge cases, enhancing the classifier's learning process. Additionally, this method eases the strenuous labeling process. AL may benefit most from this process, as it may propose unlabeled segments that are more likely to contain cues due to their proximity to labeled cues. Heilbron et al. [2018] proposed an AL framework for temporal action localization. A model is pretrained on the available labeled data. This is then used to select segments that are likely to improve the model the most. The selection function may take into account the uncertainty of a possible prediction or the likelihood of belonging to cue behavior. The research revealed the importance of sampling strategies in the effectiveness of using AL for action localization.

2.3 Machine Learning Approaches

Now that preterm infant hunger and feeding discomfort, as well as their measurement have been discussed, classification can be reviewed. Typically, supervised deep learning methods are applied for motion analysis. However, in this domain data scarcity prevents the use of traditional supervised methods. Approaches designed to circumvent this limitation must be considered. Two such approaches are Few-Shot Learning (FSL) and Active Learning (AL). FSL is an approach that exploits the existing labeled data to quickly adapt to the target domain, whereas AL queries an oracle for labels and uses an intelligent sampling strategy to minimize the labeling cost.

2.3.1 Few-Shot Learning

Few-Shot Learning (FSL) is a type of machine learning approach in which the amount of available data is only limited [Wang et al., 2020]. The goal of FSL is to train models that generalize well and generalize quickly to new targets with only a few examples available in the target classes. This approach is well-suited to the domain of infant cue classification, since only limited supervised data exists to train a model with. It eliminates the requirement for a large labeled dataset by exploiting the few samples that are available. An unspecified classification problem with N classes and K shots typically has $K \times N$ total instances in the dataset. However, this may be unbalanced between classes if more examples exist in only a specific class. Zieren and Kraiss [2005] and Kadir et al. [2004] showed that increasing the number of shots steadily improved model performance, so it is important to acquire as many *shots* as possible, until a fully supervised approach is possible.

Meta-learning algorithms aim to train a model to adapt quickly to new unseen tasks, classes, or domains with minimal examples by exposing it to a diverse range of classification problems [Jamal and Qi, 2019]. By training across these tasks, the model explores the weight space, optimizing its starting point for the target domain, which results in more efficient parameter training. This method is also referred to as task-agnostic meta-learning. Figure 2.2 shows a visualisation of the training process, split into a meta-learning phase and adaptation phase to the target domain.

Similarly, transfer learning can be applied to FSL. In transfer learning, there is a source domain and a target domain. In the source domain there exists sufficient training data to train a supervised model. The acquired knowledge can then be transferred to the target domain by retraining the final layers. This approach can be implemented to increase performance over standard supervised methods, as was shown by Gupta et al. [2020] on image classification. It is distinct from task-agnostic meta-learning, as this time the model is pretrained on another single domain, rather than a variety of tasks. Both in meta-learning and transfer learning, task relatedness should be considered when selecting the source domain [Wang and Deng, 2018]. Task relatedness concerns similarity between the two domains, where the relatedness determines if the knowledge gained in the source domain is suited to target domain. When tasks are sufficiently related, similar features may be derived, or features may be transformed some other meaningful way. Task relatedness can be defined as proximity in the feature space [Xue et al., 2007], or by similarity in features [Wilson and Cook, 2020].

Another final methodology is based on Siamese CNNs which do not learn parameters through a traditional training phase to assign an output class to an individual instance [Vinyals et al., 2016; Jadon and Srinivasan, 2021; Koch et al., 2015; Bertinetto et al., 2016]. This framework is depicted in Figure 2.3. Instead, the Siamese CNN is trained to evaluate the similarity between features of a pair of input images. This is achieved through the utilization of two identical models, each trained on the entirety of the training set. A differentiating layer is then employed to assess whether the models have learned equivalent features. To apply this model to FSL tasks, a verification and generalization stage is implemented. The verification stage requires the use of APN triplets of data instances, along with a triplet loss function. These labeled APN triplets consist of an anchor image, a positive image, and a negative image. Ideally, these images should be highly similar to train the model on hard-to-distinguish instances. The generalization phase involves training the model to distinguish whether two images belong to the same class, potentially using a different domain from the verification stage.

When an FSL problem only contains one example per class, it is called one-shot learning. The challenge of this adaptation is to learn as much information as possible from a single example in each class [O' Mahony et al., 2019], rather than a few example per class. Zero-shot learning, uses a different approach in this paradigm, with some variations in

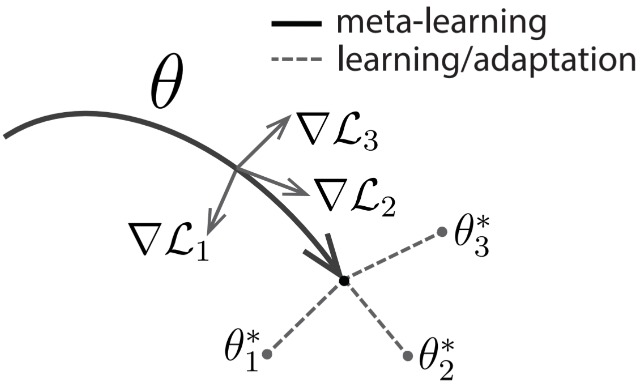


Figure 2.2: Visualisation of meta-learning, in which the meta-learning phase optimizes the weights θ given a loss function $\mathcal{L}$ to provide an improved starting point for quick adaptation to the target domain [Finn et al., 2017].

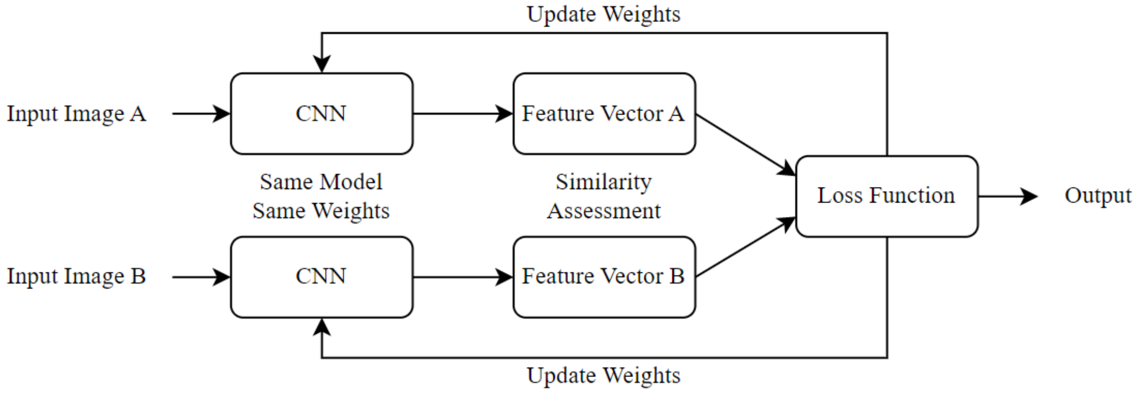


Figure 2.3: An example FSL architecture based on a Siamese CNN, during the training phase. Inputs are fed to identical CNN models. The resulting loss function is based on their output and a similarity score. In testing, a test instance is held against an anchor image.

the training and testing stage. The main objective of zero-shot learning is to recognize instances from unseen classes during testing. These classes have not been encountered during training. This method does require information from source domains, to have some kind of supervised data and allow learning [Wang et al., 2020]. To achieve this, zero-shot learning requires two sets of data, one for classes during training, and another for unseen classes during testing. The goal of zero-shot learning is then to learn a common feature space that can be used to recognize instances in unseen classes [Xian et al., 2017]. Classes exist in this feature space, and the goal of the classifier is to determine where an unseen class belongs in this space, and whether it is part of an existing or new class. K-nearest-neighbour algorithms are suited to this goal [Romera-Paredes and Torr, 2017].

After this overview of the existing approaches, it is relevant to assess how FSL has been applied in the medical field, and specifically for infant behavior analysis. Whilst there is no existing literature on infant cue classification using FSL, studies with somewhat similar target domains have been conducted. Romanov et al. [2021] did GA predictions using FSL. GA predictions are relevant because they aid in treatment decision by clinicians. They used a task-agnostic meta-learning approach, using tasks like celebrity facial recognition to train their models on a variety of tasks. Their data lends itself to a 5-way and 5-shot approach, training a handcrafted CNN. They trained on images from the GestATional Project dataset, which is a dataset consisting of images of body parts of preterm infants, like feet, hands and faces with the aim of predicting their GA. Other than this study, FSL has also been using for motion analysis, although not in preterm infants. Yang et al. [2013] used one-shot learning to recognise human actions and facial expressions using as few examples as possible. They compared supervised classification using support vector machines and unsupervised clustering. They let their model transform input videos into a feature representation using optical and then used 1-nearest-neighbour clustering to find group instances in their respective classes. Zhang et al. [2022] used FSL for autism spectrum disorder trait classification based on facial dynamics in an interview setting. They used multiple 1-hour long untrimmed videos of interviews which they had to manually segment so that it was suited for training. They extracted spatio-temporal features from the faces on participants and computed multiple different descriptors using this information. After feeding the feature vectors to their model, they found they could successfully distinguish between autism and non-autism traits. While this study was performed on adults, it does serve as evidence as to how computer vision can be used to classify states based on behavioral information using FSL.

In general, the quality of FSL is that from only few available samples, unseen classes can successfully be classified. These methods have been applied to infant data, for the estimation of their age, and to motion analysis. In this thesis, the aim is to combine these modalities, and perform motion analysis on infants. Pitfalls from the related works are that they often require some kind of available labeled dataset as point of reference or for pretraining, which is not readily available for behavioral cues.

2.3.2 Active Learning

Active Learning (AL) is a machine learning approach where the algorithm enhances its performance by actively selecting the instances from which it learns [Settles, 2009]. With this approach, acceptable model performance could be achieved with less training data. Deep learning algorithms are often trained on large volumes of data in a supervised setting. However, large labeled datasets are not available for every target domain. AL overcomes the labeled data problem by querying an oracle for the labels of selected unlabeled instances. Typically, this oracle is an expert human annotator. However, it could be any type of label provider. Figure 2.4 shows a graphical representation of this AL cycle.

So, the goal of AL is to reduce the cost of labelling, and achieve acceptable model performance [Ren et al., 2021]. The components of the approach are a machine learning model, a set of unlabeled samples, an oracle and a sampling strategy. The model can be any type of classification model, however it is assumed that it is able to learn from small amounts of labeled data. A sufficient pool of unlabeled data should already exist. If labeled samples exist in this set, they may be exploited as a starting point for the model. If they do not, then random instances may be fed to the oracle as model starting point.

When considering the sampling strategy, multiple approaches are available. It is assumed that the data is non-uniformly distributed, meaning there are instances of high uncertainty where it is more difficult to discriminate between labels and to find the correct label among

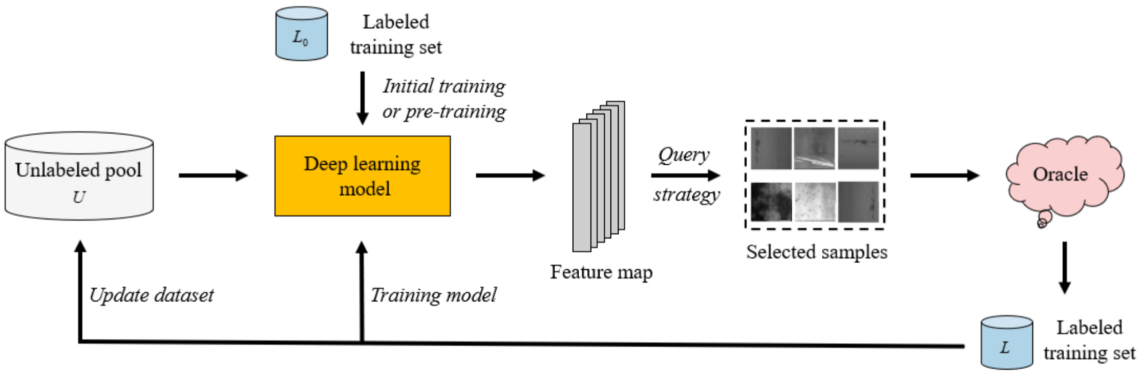


Figure 2.4: A prototype AL architecture by Ren et al. [2021]. Labeled instances are used to pretrain the model, which improve the quality of the sampling strategy. Unlabeled instances are sampled, and consequently labeled by the oracle. These now labeled instances are fed to the model again and removed from the unlabeled pool, from which new examples are selected in a new iteration.

a set of at least two outcomes. The most popular approach is an uncertainty-based approach [Settles, 2009], in which the samples of the highest uncertainty are selected. The idea is that the most uncertain instances border the class labels and therefore have discriminative value. This approach may also be adapted to use entropy or feature diversity rather than uncertainty [Holub et al., 2008]. Another approach is to use the lowest difference in probability between the most likely and second most likely class as uncertainty score (best versus second-best), which is more greedy. Despite this, it was empirically proven to compete with state-of-the-art sampling strategies [He et al., 2022].

With all components in place, related works are discussed. Due to the novelty of an AL approach, very few published works implementing this approach for videos of infant behaviors exist. However, AL has been successfully used in the medical field. Budd et al. [2021] reviewed studies that work with medical image analysis, with a human annotator as oracle. They concluded that the adaptions of AL, and other deep learning techniques for that matter, may cause a paradigm shift for many clinical tasks, from assessment to treatment. Another possible implementation of AL is to use the trained model as annotator for future data, where medical experts may not have the time to do so. However, Lowell et al. [2018] raised the concern that datasets may at some point outgrow the trained model, and performance steadily decreases so that it serves no practical application. Therefore the generality of the model must be kept in mind when assessing its quality, and it must be robust for natural variation in the training data that possibly does not occur in the training set. A final relevant take on AL uses transfer learning. In a situation where the target domain does not hold sufficient data for successful training, one may seek out data from a different but similar domain to improve training. When using labeled external data, this has already proven to be successful [Pan and Yang, 2010; Zhu et al., 2011], as well as in the medical setting where training data is typically scarce [Ravishankar et al., 2016]. In the context of AL, an external domain may contain many unlabeled instances, but may improve learning in the target domain. In their study, Zhu et al. [2011] propose a method that starts with several labeled data in the target domain and to iteratively label data from the external domain. During a training iteration, the informativeness or weights of the external dataset are updated based on the prediction errors of the classifiers. This method achieves better accuracy and convergence speed than traditional methods, and may also add generality to the model. This addresses the previous concern raised by Lowell et al. [2018] that datasets may outgrow models.

Settles [2009] in their literature survey confirmed that AL approaches with uncertainty sampling can outperform standard supervised learning and other AL strategies like the diversity-based approach or approaches that focus on the cost of labeling instances and random batch sampling. As demonstrated by Lorbach et al. [2019] in their study, AL can be particularly useful in annotating the more chaotic and less goal-oriented behavior of rodents. This is similar to the behavior of infants, who also may not be able to effectively communicate their needs and show large variation in their behaviors. However, Houlsby et al. [2011] noted that the effectiveness of uncertainty sampling is context-dependent and is further determined by the quality of the sampling metrics. This may be relevant to the context of this study, as AL approaches are not well-researched in the NICU setting. Since cue-based behavior may show individual differences impacting classification, it is possible that the model may require a new learning phase for each new infant. Ideally, this would not be necessary as the weights from previous iterations should generalize well to new data. However, AL is well-suited for this process, as it allows for the model to undergo additional learning iterations.

3 Methodology

The upcoming section details the methods used in this thesis. It starts with an account of the datasets, collection, annotation, preprocessing, feature generation and train-test splits. Next, the cues that are suitable for automated detection are discussed. The section continues by reviewing the MoViNet models and their architecture. Finally, evaluation metrics, hyperparameter tuning and software used for model training are discussed.

3.1 Data

The SLAPI dataset was used in this study to analyze infant behavioral cues, including sleep, hunger, and feeding discomfort cues. The SLAPI dataset contains videos of preterm infants admitted to the NICU of the UMC Utrecht, and videos were recorded of preterm infants of any GA. Only authorized individuals could access the data, which was securely stored on the UMC Utrecht servers. While the primary goal of this dataset is to assess sleep, the recorded data is also suited to the classification of behavioral cues. However, due to the dataset's developmental status, the data available for analysis in this study was sparse. Due to this sparsity, the cues considered in the experiment must be carefully selected for their fitness and availability.

The carefully chosen pool of cues, as expanded upon in Section 3.2, is necessary to train the baseline, FSL and AL models. These first two require sufficient data for a fully supervised approach, while the AL model can be used to expand upon the available samples in the user study. The dataset was discovered to be imbalanced, as some behaviors are more frequent in some infants across their states, than in others. Table 3.1 shows the class distribution by frequency, duration and size. The classes "*Still*" and "*Arm move*" have relatively many instances with 149 and 97 instances respectively, while classes like "*Rooting*", "*Shiver*", "*Tensing hand*" and "*Tugging*" are underrepresented with less than 10 instances. As for duration, classes like "*Still*" and "*Crying*" stand out with 16 seconds per snippet on average, while "*Eye squeeze*", "*Shiver*" and "*Tongue out*" are at most 4 seconds.

For model pretraining and meta-learning, external datasets were used. It is essential that external datasets contain as many outcome classes as possible, with the goal of training the model to classify based on as many intricacies in the data as possible. These models can then generate features that are sufficiently detailed for subtle movements in infants, as well as robust enough for variations within classes. Kondratyuk et al. [2021] used the Kinetics600 (Section 3.1.7) dataset for pretraining the MoViNet models, and the UCF101 (Section 3.1.8) dataset was selected for the meta-learning tasks.

3.1.1 Participants

Preterm infants with a GA of at least 32 weeks were included in this study. Preterm infants with a GA of at least 32 weeks start to display behavioral indicators that indicate their state, as Whetten [2016] noted in their research. So, infants with a GA of at least 32 weeks were chosen in order to ensure that they exhibit these behavioral cues.

Participant recruitment was carried out in coordination with the neonatology department of the UMC Utrecht. When a preterm newborn met the inclusion requirements, the

Category	**Class**	**Duration (*s*)**		**Samples**			
		M	*SD*	*Train*	*Val.*	*Test*	*Total*
Other	Adult hand	12	8	31	15	17	63
Other	Arm move	7	4	48	24	25	97
Hunger	Bottle feeding	11	8	13	6	7	26
Hunger	Bottle sucking	14	7	15	7	9	31
Discomfort	Crying	16	47	27	13	15	55
Discomfort	Eye squeeze	4	1	9	4	6	19
Discomfort	Finger splay	11	7	5	2	4	11
Discomfort	Flexing	8	4	5	2	3	10
Hunger	Fussy	10	5	18	9	9	36
Other	Grasping	7	4	5	2	4	11
Other	Hand move	10	9	31	15	17	63
Hunger	Hand sucking	8	3	11	5	6	22
Hunger	Hand-to-face	8	6	26	13	13	52
Hunger	Hand-to-mouth	8	5	24	12	12	48
Other	Head move	11	13	8	4	5	17
Hunger	Mouth open	5	3	20	10	11	41
Hunger	Mouthing	8	4	6	3	3	12
Hunger	Rooting	6	4	4	2	2	8
Other	Shiver	3	1	4	2	2	8
Other	Still	16	18	74	37	38	149
Discomfort	Tensing hand	7	2	2	1	2	5
Hunger	Tongue out	4	3	14	7	8	29
Other	Tugging	5	2	4	2	2	8
Other	Yawn	5	1	13	6	8	27
Mean		8.5	—	—	—	—	—
Total		—	—	417	203	228	848

Table 3.1: Overview of class distribution in the SLAPI dataset, including tentative division of classes into infant state categories.

Infant ID	**GA**	**Sex**	**Video Length**	**Total Frames**	**Unique Cues**
50	37.4	M	1:36:47	176004	18
70	35.1	F	0:30:38	55711	19
79	35.3	M	1:00:55	110777	21
95	35.7	F	0:58:52	107059	19
74	35.4	M	1:27:40	157814	9

Table 3.2: Demographic details and video-specific information for each infant included in the SLAPI dataset. GA: Gestational Age (weeks). Video Length: Hours:Minutes:Seconds. Total Frames: Number of frames in the video. Unique Cues: Number of distinct cues observed.

parents were contacted and informed about the study. Parents received comprehensive information of the study's purpose, the video recording process, and the advantages and disadvantages of taking part. Parents who agreed to their child taking part in the research gave written informed consent. Ethical approval was granted to the SLAPI study to create recordings.

In the end, the recordings of 5 infants were included in the dataset used for cue classification. These recordings were selected based on the display of cue behavior, the visibility of the infants' faces and upper bodies and the absence of occlusions. The untrimmed videos of infants 50, 70, 79, 95 and 74 were annotated and included in the SLAPI dataset. Table 3.2 shows infant-specific information pertaining their videos and demographics.

3.1.2 Collection

Infant safety was prioritized during the data collection. Data was collected under regular NICU circumstances, without any disturbances to care interventions, parental visits or feeding. The cameras were placed on the top and side of the NICU bed, allowing easy access for nurses. The videos were recorded when parents were not present, and stopped whenever necessary. Upon completion of a recording, the video pseudonymized and safely stored on the servers of UMC Utrecht.

The data collection was conducted using the VDO360 webcam camera. It records video at 1920x1080 resolution with 30 frames per second using a 2-megapixel sensor. The camera's field of view is 70 degrees, which enables sufficiently wide coverage of the NICU bed. Its dimensions are 38x38x44.6mm making it compact enough for the NICU, without obstructing medical care.

3.1.3 Annotation

Snippet annotation was carried out by research interns at the UMC Utrecht, including the author of this thesis. They did not have a medical background other than their internship. The annotators carefully watched the untrimmed videos and documented the start time and duration of any behavior by the infant. These behaviors could include all but the physiological cues from Table 2.1, Table 2.2 and Table 2.3. This process resulted in approximately 150 snippets per untrimmed video, recorded in an excel sheet from which a Python program could extract the snippets. The quality of the annotations is crucial to the quality of any model trained on this dataset. Uniformity of the labeling was ensured by discussion between the annotators, so that ambiguous or simultaneous cues were labeled in similar fashion.

The AL study (see Section 4.3) served a dual purpose, by not only functioning as experiment, but also as data annotation tool. Previously unlabeled snippets were evaluated and labeled by an oracle. All those annotations were stored and incorporated in the larger labeled dataset. These annotations were not included in the baseline and FSL studies to maintain fair comparison between the experiments and their benefits or drawbacks. Despite their exclusion from these experiments, they were included in the final dataset, for use in potential future analyses. These snippets were also included in one ablation condition, to determine the impact of larger datasets on performance.

3.1.4 Preprocessing

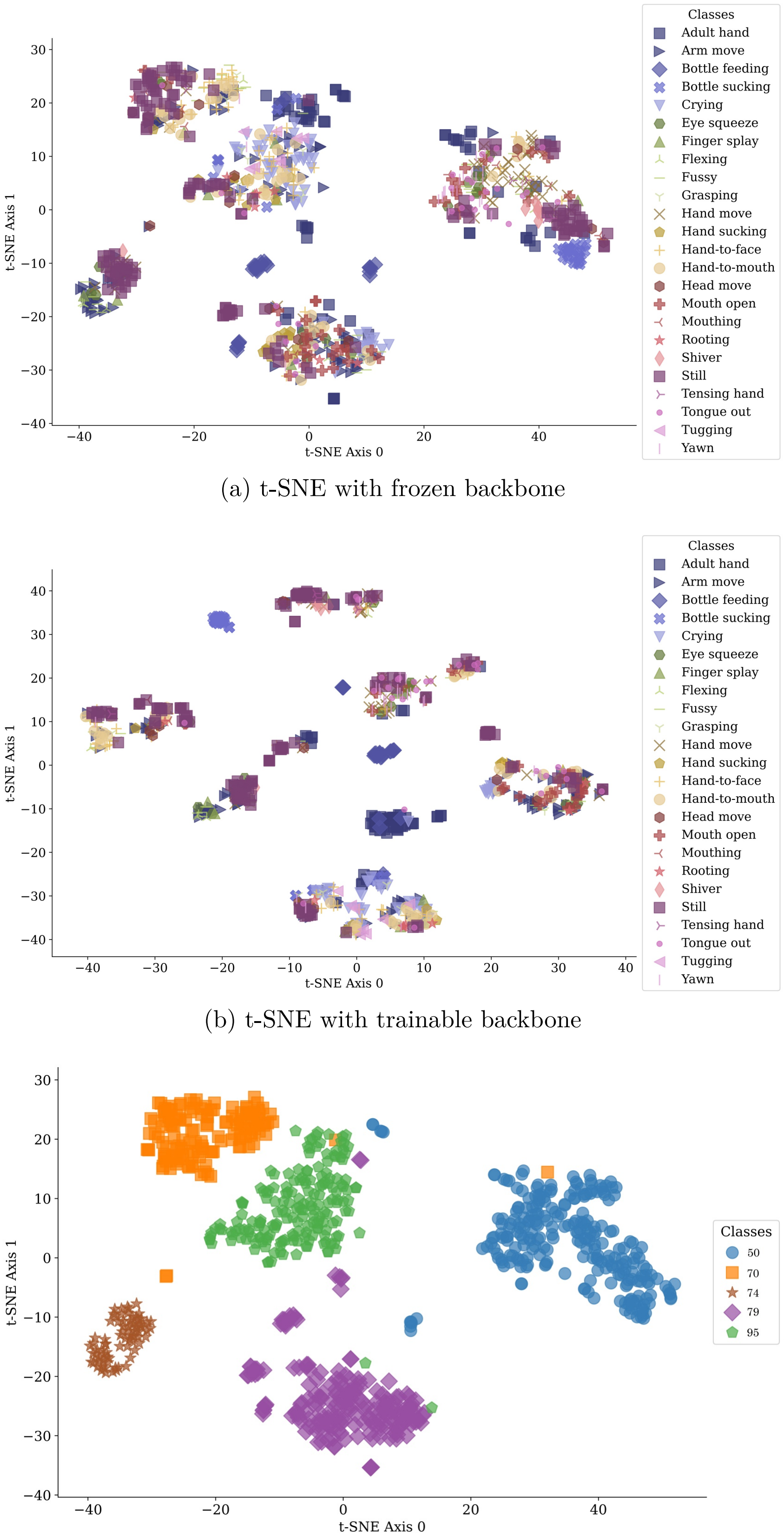
The preprocessing of the SLAPI dataset, which consists of hour-long untrimmed videos, was an important step to adjust it for the snippet-based models. First, videos were rotated to ensure all the infant's heads were pointed upwards. This was followed by cropping the video frame, with the goal of decreasing the size of the video and hereby maximizing the information retained when the resolution was reduced to fit the model's preferred resolution. Cropping was done manually by the annotator after determining the position of the infant across the video. Resizing was done using bilinear interpolation, maintaining the aspect ratio of the original frames. The result was padded with zeroes to fit the target shape.

After these changes, the untrimmed videos were segmented into labeled snippets, applying the earlier annotations. These snippets were then processed further to form the final dataset. This includes extracting equidistant frames from each snippet. The number of frames used in the representation is a hyperparameter of the MoViNet model. If the video did not contain sufficient frames, empty frames were appended to the back of the representation to conform to the required shape for the model inputs. Additionally, the resolution of the videos was reduced to align with the preferred resolution of the MoViNet model used. The outcome of this processing was a dataset with shape (batches × frames × width × height × 3).

For the AL generated proposal snippets, the objective was to create snippets of 5 to 15 seconds, corresponding with the preferred number of FPS by the model. This algorithm for proposal generation is explained in Algorithm 3. The main difference in preprocessing lies with the selected frames. As consecutive clips with the same label are added together to generate longer snippets, the number of frames is halved to maintain shape consistency. An additional benefit of combining clips this way, is the actual FPS in the processed snippets are somewhat similar to existing datasets, and deviate somewhat from the preferred FPS which is inevitable due to varying snippet lengths. The snippets were not subjected to further preprocessing. The motivation behind this approach is that the model is exposed to the rawest form of the data, so that it can learn the natural variations in the data. However, exploring other preprocessing techniques makes an interesting ablation condition, to gain insight into how it may have affected performance. It has been shown that more extensive preprocessing can improve model robustness by enhancing the model's understanding of temporal patterns [Rebuffi et al., 2021]. However, this decision should be weighted against additional complexity and computational cost.

3.1.5 Feature Generation

To create a feature representation of the snippets, all model layers but the final classification head were used as feature generator. Specifically, the penultimate layer of the CNN is utilized to extract high-level features from the input videos. This layer has been shown to capture semantically rich features that are useful for a wide range of tasks both in the spatial and temporal domain. By using the penultimate layer for feature extraction, the need for explicit feature generation was avoided. This approach relies on the CNN's ability to learn discriminative features from the data, making it more robust and adaptable. Snippets were represented in a one-dimensional array with 640 features, encoding both spatial and temporal information.



(a) t-SNE with frozen backbone

(b) t-SNE with trainable backbone

(c) t-SNE labeled by infant

Figure 3.1: t-SNE visualization of the SLAPI dataset using MoViNet A2 model. Each marker denotes a video snippet, separated by marker style and color based on class. Proximity indicates feature similarity in a reduced 2D space. (a) uses the model as-is, (b) shows fine-tuned model results, and (c) is labeled by infant source video.

The features generated by the MoViNet A2 model are shown in Figure 3.1. The dimensionality of the data is reduced to visualise the distribution using the t-SNE technique [Van der Maaten and Hinton, 2008]. While maintaining the local structure, it uncovers global patterns like clusters, offering insights into the distribution of the dataset that help interpret model performance. When the model backbone is fine-tuned on the dataset, as shown in Figure 3.1(b), some cues are well-clustered. "*Bottle sucking*" and "*Adult hand*", for example hardly overlap with other classes. The t-SNE representation also managed to somewhat single out "*Hand-to-face*" despite some overlap. "*Still*", on the other hand, overlaps with many classes. In Figure 3.1(a) the backbone is not fine-tuned. As a result, the classes overlap more. However, there are still clusters of classes that are clearly distinguishable, like "*Adult hand*", "*Bottle sucking*" and "*Bottle feeding*". In terms of inter-class variance, it can be appears that the variance is not sufficiently large to clearly distinguish between classes by visual inspection. Intra-class variance, on the other hand, is substantial, with smaller clusters of the same class appearing across the 2D plot. The dataset includes five untrimmed videos, and it appears that the visualization reveals five individual clusters, each potentially belonging to a specific video. Figure 3.1(c) was created to visually evaluate that the these clusters belong to each infant. Rather than coloring the instances by class, the t-SNE plot is colored per source video. This confirms the idea that the clusters belong to the source videos, meaning the feature representation does not only represent the target classes, but also carries distinct signatures of the originating videos. Ideally this would not occur.

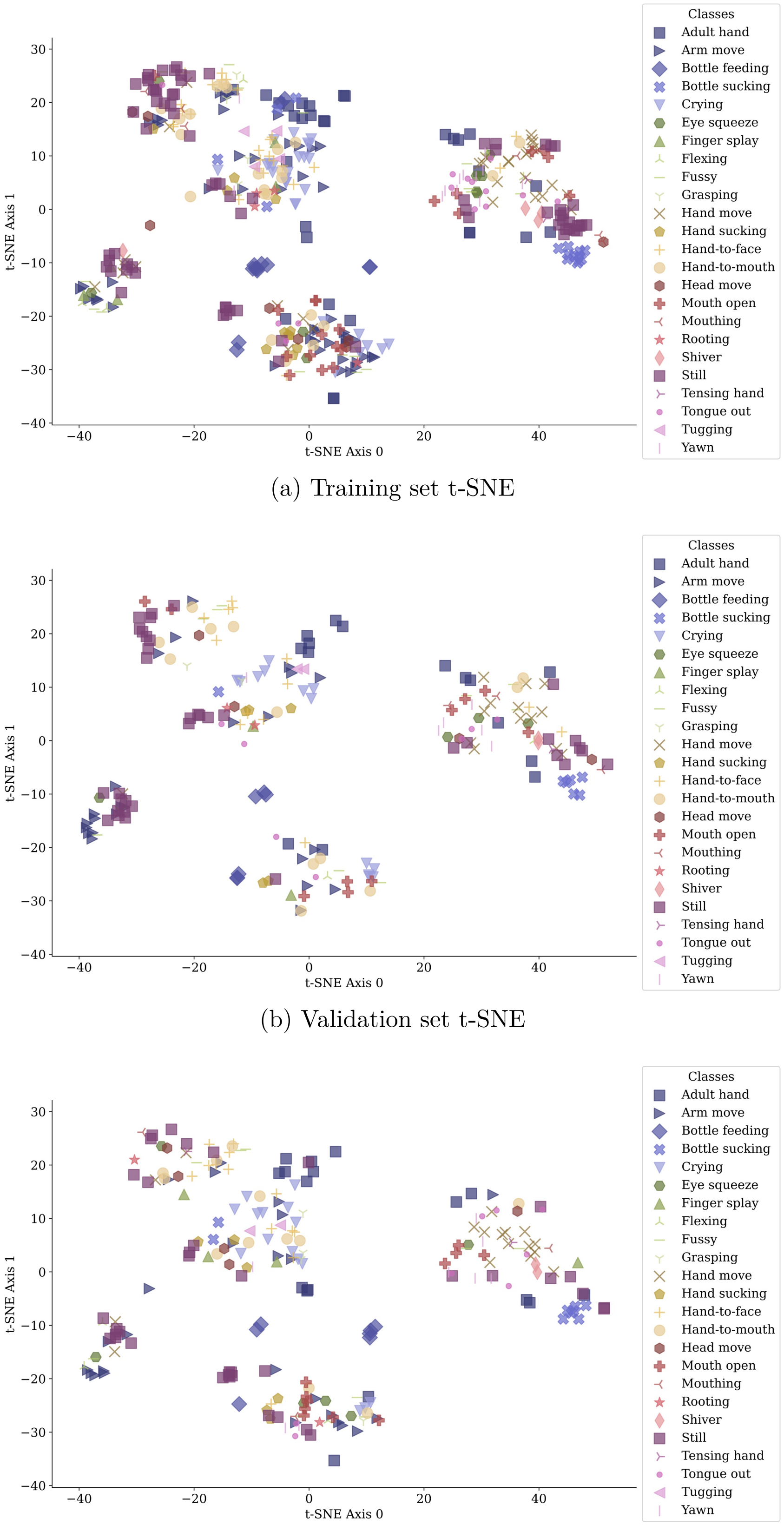
3.1.6 Train, Validation & Test Split

The data was split using a train ratio of .50, validation ratio of .25 and test ratio of .25. Snippets from each infant were distributed over all different sets, if there were sufficient snippets to do so. The test set was allocated any remaining snippets. The split configuration that was used in all experiment is documented in the GitHub repository (https://github.com/joris-s/cues). It is recommended that any replication efforts use this configuration. The t-SNE representations per set are shown in Figure 3.2, which may help interpret discrepancies between train, validation and test accuracy. The consistency between figures shows that the train, validation and test sets have comparable underlying structures, leading to a consistent representation of the data throughout model evaluation.

Due to the class imbalance, the number of samples was limited to 20 samples per class in the training set. This way, the maximum discrepancy between classes is 16 instances. The split was kept consistent for all experiments throughout this thesis, except for the component in the ablation study (see Section 4.4) where the model is tested on multiple splits on the data. It uses entirely different splits which were randomly generated.

3.1.7 Kinetics600

Kay et al. [2017] developed the Kinetics human action video dataset, which contains 400, 600 or 700 classes of human action in roughly 10 second videos scraped from YouTube. The dataset contains between 400 and 1150 clips for each action, from unique videos. It was developed as successor to the UCF101 dataset. When the dataset was released, it originally consisted of 400 classes, but has now been expanded with variations of 600 and 700 classes. The dataset has been widely used for (pre)training action recognition or



(a) Training set t-SNE

(b) Validation set t-SNE

(c) Test set t-SNE

Figure 3.2: t-SNE visualisation of the SLAPI dataset with data points separated by their respective train, validation, and test splits. The consistency between the figures shows that the train, validation and test sets have comparable underlying structures, leading to a consistent representation of the data throughout model evaluation.

localization models [Clark et al., 2019; Du et al., 2021; He et al., 2019; Cherian and Gould, 2019], and has been shown to significantly improve models when used as pretraining dataset.

The videos are realistic amateur videos with differing quality, camera stability and lighting. This makes classification challenging, due the the great intra-class variance. However, this also makes the challenge representative of real-world scenarios, to the benefit of the generalization qualities of models. The dataset contains different categories of classes, like singular person actions, person-person actions and person-object actions.

The MoViNet architectures used in this thesis were all pretrained on the Kinetics600 dataset, allowing for a feature transformation that is sensitive to details due to its many subtle differences between classes.

3.1.8 UCF101

Soomro et al. [2012] created the UCF101 dataset, which was widely used for action recognition. It consists of roughly 13.000 clips in 101 action classes. These classes are categorized into five categories of actions: human-object interaction, body-motion only, human-human interaction, playing musical instruments and sports. A unique feature of this dataset is its high intra-class variability, due to the diversity in environment, lighting conditions and camera quality and viewpoints. It was used state-of-the-art performance in large scale action recognition tasks [Karpathy et al., 2014; Saha et al., 2016; Weinzaepfel et al., 2015].

Due to its high degree of intra-class variability, it a suitable dataset for meta-learning tasks. The goal of meta-learning is to train a model to adapt its weights to new, unseen environments or classes. The diversity of data provides a rich source for the algorithm to learn to generalize. It is therefore selected for the meta-learning tasks for the FSL model described in Section 4.2.

3.2 Cues

The objective of this thesis is to classify hunger and discomfort behaviors in an effort to detect hunger and feeding discomfort, using cues from the SLAPI dataset. The classes were tentatively divided into three categories: "*Hunger*", "*Discomfort*" and "*Other*". Due to the ambiguous nature of infant cues, as discussed in Section 2.1, they may belong to multiple categories, and therefore the categorization proposed in this section is not definitive. For instance, the behavior "*Arm move*" may be an indication of hunger when the infant is rooting or this movement is towards the mouth. It could also function as self-soothing or as a random movement during the active sleep phase.

However, this categorization is grounded in background literature (Table 2.1 and Table 2.2), as well as the insights gained from discussions with three NICU nurses. They indicated that hunger cues largely center around the mouth, in the form of licking lips, mouthing, tongue tauting and sucking behaviors. When a pacifier or food is provided to the infant and it calms down immediately, it is generally interpreted as a clear sign that the behaviors were related to hunger. As for discomfort cues, those behaviors were displayed on the entire body, like creating a fist, arching or flexing, and frowning.

So, in the SLAPI dataset, hunger cues are "*Bottle feeding*", "*Bottle sucking*", "*Fussy*", "*Hand sucking*", "*Hand-to-face*", "*Hand-to-mouth*", "*Mouth open*", "*Mouthing*", "*Rooting*" and "*Tongue out*".

Discomfort cues are "*Crying*", "*Eye squeeze*", "*Finger splay*", "*Flexing*" and "*Tensing hand*".

Other than the hunger and discomfort cues, there are multiple behavior that do not belong to either of these main categories. They were all grouped together and are "*Adult hand*", "*Arm move*", "*Grasping*", "*Hand move*", "*Head move*", "*Shiver*", "*Still*", "*Tugging*" and "*Yawn*".

3.3 MoViNet Models

Kondratyuk et al. [2021] developed MoViNet (short for Mobile Video Network) neural networks for efficient (online) classification of videos. Standard 3D CNNs are effective at video recognition, however, they are computationally heavy and have high space and time complexity. To address these limitations, they propose an approach with three steps that includes a neural architecture search, a causal stream buffer technique to decouple memory from video duration and an ensembling technique to further improve accuracy. They constructed multiple networks, with different decisions made in the accuracy versus complexity trade-off. Their networks range from MoViNet-A0 to MoViNet-A6 with base, streaming and ensembling variants. These networks differ in their structure, and as a result the number of parameters and GFLOPS for classification. These different models were found by placing constraints on the architecture search. This search starts from a MobileNetV3 which was previously used successfully for video classification [Koonce, 2021], and searches for all hyperparameters, ranging from input dimensions, frames per second, kernel size, filters and number of blocks. Note that this is not an exhaustive list due to the complexity of the search. The result is a family of MoViNet models which can be used for video classification. These models are inherently suited to the cue classification problem as it is a video snippet classification problem.

The MoViNet architecture consists of three key components. First, depthwise separable convolutions convert a standard convolution into a depthwise and point-wise convolution. The depthwise convolution applies separate filters for each input channel, and the point-wise convolution uses a `1x1` convolution. This significantly reduces memory and computational complexity. In small networks, however, this reduction in parameters may lead to the model being unable to properly construct features from its input. Second, squeeze-excitation blocks transform the feature maps by (1) multiplying the map (excitation) with a constant derived from the maps using average pooling, and (2) fully connected layers summed to a single value after smoothing (squeeze). This way, a global understanding of the feature maps is incorporated in the features. Third, causal convolutions replace all temporal convolutions making them unidirectional, and unable to violate the ordering of the frames.

The A0 model is the least complex and fastest model, with 3.1 million parameters. This rises to 31.4 million parameters for the A6 variant. Only the A3 model breaks the trend, as it has more parameters than both the A2 and A4 model, with 5.3 million parameters. This parameter space is still significantly smaller compared to models like SlowFast-R152 and EfficientNet-L2 with 80 and 480 million parameters respectively [Kondratyuk et al.,

Model	Resolution	FPS	Parameters	GFLOPS
A0	172x172x3	5	3.1M	2.71
A1	172x172x3	5	4.6M	6.02
A2	224x224x3	5	4.8M	10.3
A3	256x256x3	12	5.3M	56.9
A4	290x290x3	8	4.9M	105
A5	320x320x3	12	15.7M	281
A6	320x320x3	12	31.4M	386

Table 3.3: Comparison of Movinet Models. Resolution indicates the width and height of the input, FPS the preferable frames per second of the video. Complexity can be inferred from the model's parameter count and GFLOPS [Kondratyuk et al., 2021].

2021]. The models require increasing computations as they become more complex. Top-1 accuracy on the Kinetics600 dataset increases with model complexity. For the A0, A1 and A2 model, the base variant without streaming buffers or ensembling enhancements provided the best top-1 accuracy on the Kinetics600 dataset. The streaming-ensemble variant performed best on the Kinetics600 for the A3, A4 and A5 models. The A6 model was not tested on the Kinetics600 dataset. Table 3.3 gives an overview of the models and their differences, including preferred hyperparameters and architectural differences.

In this study, only the A2 model was used due to its comparable number of parameters with models A0-A4, as well as the low number of GFLOPS required for testing. Kondratyuk et al. [2021] also found the A2 model offered an acceptable balance between performance and computational cost. It scored the third best top-1 accuracy on the Kinetics600 dataset, only surpassed by the A4 and A5 models.

3.4 Evaluation Metrics

To evaluate the model performance on cue detection and classification, the accuracy, precision, recall and F_1 score are recorded. Table 3.4 gives an overview of these metrics and how they can be computed.

Accuracy measures the proportion of correctly classified examples among all examples. It is computed by finding the ratio of true positives and true negatives to the sum of true positives, true negatives, false positives, and false negatives. While accuracy is a simple metric for overall performance, it does not provide information about performance on specific classes, and is less informative when class imbalance occurs in the dataset. However, it is still included as metric to give a general indication of model performance. Often, balanced accuracy is computed as the average of recall scores across classes. This way, more importance is given to the underrepresented classes.

Precision measures the proportion of true positives among all examples that are classified as positive by the classifier. It is computed by finding the ratio of true positives to the sum of true positives and false positives. Lower precision scores indicate that the model classifies too many false examples as part of the target class.

Recall measures the proportion of true positives that are correctly identified by the classifier. It is calculated as the ratio of true positives to the sum of true positives and false negatives. Lower recall scores indicate that positive examples are not identified correctly.

Metric	Formula	Explanation
Accuracy	$\frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i+FP_i+TN_i+FN_i)}$	Proportion of correct predictions of total predictions.
Precision	$\frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i+FP_i}$	Average proportion of true positives of predicted positives.
Recall	$\frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i+FN_i}$	Average proportion of true positives of actual positives.
F_1	$\frac{1}{C}\sum_{i=1}^{C} 2\frac{Precision_i \times Recall_i}{Precision_i+Recall_i}$	Average of the harmonic mean of P and R of each class

Table 3.4: Evaluation metrics used in this study to evaluate model performance for cue classification. TP stands for True Positive, FP for False Positive, TN for True Negative, and FN for False Negative. i is the class instance and C represents the total number of classes. The metrics are macro-averaged, averaging every class equally.

Finally, the F_1 score is commonly used when dealing with imbalanced classes. It represents the harmonic mean of the precision and recall scores. Precision and recall may be a trade-off between each other. When trying to identify the most positive instances, that means that negative examples may be classified as positive in the process. Therefore, a higher F_1 score indicates a good balance between precision and recall, and in the unbalanced setting it prevents favouring classes that are more numerous.

In the context of fairness, precision and recall can help identify whether a model is treating classes differently. When precision and recall scores are equal among classes, that means a model treats those classes fairly. It does not over- or under-predict positives and is not disproportionally failing to find positive cases. F_1 then provides a more comprehensive score by combining precision and recall.

3.5 Hyperparameter Tuning

The MoViNet architecture is defined by a large set of predetermined hyperparameters. As a result, typical CNN hyperparameters such as the number of layers, units, filters, and kernel size, do not require further refinement. However, there are remaining hyperparameters related to the training of the model that require hyperparameter tuning. These parameters are the dropout rate, the number of frames, batch size, epochs, learning rate, regularization method, enabling model backbone training and applying causal convolutions. For FSL the number of meta-tasks and shots are included as well.

Grid search is a standard approach to hyperparameter tuning. This approach tests all possible combinations of parameters and their fitness is evaluated. However, due to the long training times of MoViNet models, an exhaustive grid search was not feasible. As a result, alternative methods such as Bayesian optimization and genetic algorithms, were explored. Alibrahim and Ludwig [2021] found that genetic algorithms outperformed both grid search and Bayesian optimization in terms of time consumption. All three approaches resulted in similar metrics for search and model performance, although the model architectures varied across each approach. Based on these findings, the genetic search algorithm was selected as the most feasible method for this thesis. The implementation of the genetic search algorithm is shown in Algorithm 1 (Appendix A), with selection of individuals done through tournament selection. During this hyperparameter optimization process, the SLAPI dataset was divided into the same main training, validation, and test subsets. Consequently, the hyperparameters identified through this genetic search algorithm may only be suited to this dataset configuration. This creates a interesting ablation condition where the dataset is repeatedly reinitialized to analyze the impact of different data shuffles on performance.

The genetic search approach itself has three hyperparameters, being population size, number of generations, and tournament size. Vrajitoru [2000] concluded that using a larger population size resulted in better performing models than increasing the number of generations did. Consequently, a population size of 10 with 5 generations was chosen for this thesis. The tournament size was set to the default value of 3. Due to memory and computational constraints, a more extensive genetic search could not be conducted.

3.6 Software

The experiments were built using Python as programming language. Crucial packages for the development of the experiments were TensorFlow official models for the MoViNet implementation, TensorFlow for the remaining neural network architecture requirements and data management, and OpenCV for image processing and the user interface of the user study. When using `pip install package`, the required dependencies are also installed. The exhaustive list of the packages used for the experiments can be found in the GitHub repository (https://github.com/joris-s/cues). The experiments were conducted on the high performance computing facility of the UMC Utrecht, where the SLAPI dataset could be safely loaded.

4 Experiments

This section presents the setup for the three main experiments of this study. First, the baseline experiment is introduced, and implementation details are discussed. Then, the FSL experiment is discussed, including the meta-tasks and classes. Finally, the AL experiment is introduced, covering the user study, sampling strategy and candidate generation.

The primary objective of these experiments is to gain insight in how cues can be detected with machine learning, and what the pitfalls are. Expected outcomes of these exploratory experiments concern which cues can be effectively classified, which approaches perform best, as well as what can be expected from the models when more annotated data is added. To this end, the MoViNet A2 [Kondratyuk et al., 2021] network that was originally proposed for action classification has been adapted for cue classification in infants. The baseline experiment aims to determine the minimum expected performance and gain a deeper understanding of the advantages provided by FSL and AL. Where these methods differ primarily is in the training phase, and the MoViNet models have been adapted accordingly.

For the models to be successfully implemented, the model should be able to (1) classify video segments to their cue classes, that have been preprocessed to snippets for training and testing, (2) detect cue classes from a live video feed, frame-by-frame or in groups of frames for actual use, and (3) find cues in a longer untrimmed recording of the infant for the AL experiment.

Outside of the fixed MoViNet hyperparameter selections, the hyperparameters are summarized in Table 4.1. These hyperparameters were derived from a genetic algorithm search, as described in Section 3.5. Approach-specific parameters are discussed in their respective subsections.

Lastly, due to the many different possible architectures of the models, an ablation study was conducted to determine the importance of each component to the final models. In the experimental approaches, regularization, model backbone training and causal convolutions were not selected. Regularization helps prevent overfitting, a significant risk with the large MoViNet models. Causal convolutions, on the other hand, are used due to their their efficiency within the MoViNet architecture. It was expected that such techniques would yield advantageous results for a computationally expensive task such as video classification. Similarly, it was thought that model backbone training would have fine-tuned the feature representation to benefit classification. From the t-SNE representations of the dataset it arose that instances are somewhat clustered by class, but also by their untrimmed source video. Therefore, the absence of these elements in the experimental approaches provides interesting ablation conditions.

Approach	DO	#Frames	BS	Epochs	LR	Reg.	Backbone	Causal conv.
Baseline	0.3	14	4	5	1e-3	None	Frozen	Disabled
FSL	0.5	24	8	5	1e-2	None	Frozen	Disabled
AL	0.3	14	8	12	1e-3	None	Frozen	Disabled

Table 4.1: Summary of hyperparameters used in the different experimental approaches: baseline, FSL, and AL. The table displays variations in dropout rate (DO), number of frames (#Frames), batch size (BS), epochs, learning rate (LR), regularization, backbone training and use of causal convolutions.

4.1 Supervised Baseline Experiment

Due to the innovative application of a machine learning approach to infant monitoring, there is no comparable baseline performance available. Therefore, a baseline experiment was included in this study. The results from this experiment will serve as a point of comparison with the FSL and AL approaches, allowing for a clearer interpretation of their performance improvements or limitations. This baseline is subject to the same data limitations as the other approaches, without their architectural advantages. Therefore similar or improved peformance is expected in the subsequent experiments.

4.1.1 Framework

A standard motion classification framework, like shown in Figure 2.1, was employed for the baseline experiments. This included data acquisition by the SLAPI study, as well as motion capture and feature extraction by the MoViNet A2 model. Up to and including the penultimate layer of the network, it acts as a feature generator for the final classification layer. These features are robust due to the pretraining of the model using the Kinetics600 dataset. After training, the model is then evaluated based on the metrics accuracy, precision, recall and F_1 score based on performance on the test set.

4.1.2 Hyperparameters

The genetic search algorithm was employed to determine the parameters for dropout, number of frames, batch size, learning rate, regularization, the training of the backbone, and causal convolutions. The following parameter settings were established: dropout at 0.3, 14 frames per snippet, batch size of 4, learning rate at 1e-3, no kernel regularization, freezing the backbone, and disabling causal convolutions. Notably, the individuals in the population exhibited consistently better performance with a batch size of 4, despite research by Keskar et al. [2017] suggesting that smaller batch sizes lead to noisier weight updates which may harm convergence and therefore performance.

The loss function that was used is sparse categorical cross-entropy, which is the standard for multi-class classification problems. The optimizer used was Adam, which excels in its efficiency and adaptability to noisy gradients [Kingma and Ba, 2017]. With just 5 epochs, the training phase provided optimal results. Validation loss decreased after four or five epochs during model construction and provisional testing, and further training resulted in overfitting. Given the pretraining, this brief training period was sufficient.

To prevent overfitting during the training process, or simply learn the distribution of the data, class weighting was applied to the loss function. This ensures that samples in the underrepresented classes are weighted similarly to overrepresented classes, avoiding a bias towards such classes.

4.2 Few-Shot Learning Experiment

Building upon the baseline experiment, this section introduces the FSL experiment. This experiment was designed to evaluate the effectiveness of the enhancements that this approach brings to infant monitoring when only limited labeled data is available, benefiting from training on an external dataset. This experiment consists of two stages, being meta-learning and meta-testing. The meta-learning phase will train the model to quickly adapt its parameters to new unseen tasks. The meta-testing phase then aims to exploit this by quickly adapting model weights to the limited SLAPI data that is available. The results from these experimental runs will be compared with the baseline and AL experiments to provide an understanding of the improvements and limitations of this method.

4.2.1 Framework

This approach expands upon the standard motion analysis framework laid out in Section 4.1.1. Figure 4.1 shows a visualization of the approach. In the meta-learning phase, there is a finite number of episodes consisting of a similar classification problem with the same N classes and K instances per class in the training set. The validation loss is computed using Q query images in the validation set. These may differ in dimensions from meta-testing since the model was not trained on this set.

Then, in meta-testing, the model is trained for the target domain. This includes the train, validation and test sets of the SLAPI dataset. To allow direct comparison with the other two approaches, the model is again evaluated based on accuracy, precision, recall and F_1 score on the test set. The meta-learning phase prepared the model for swift parameter changes to the target domain in meta-testing. As more samples are collected in this approach, it progressively converges toward the standard baseline approach to video classification.

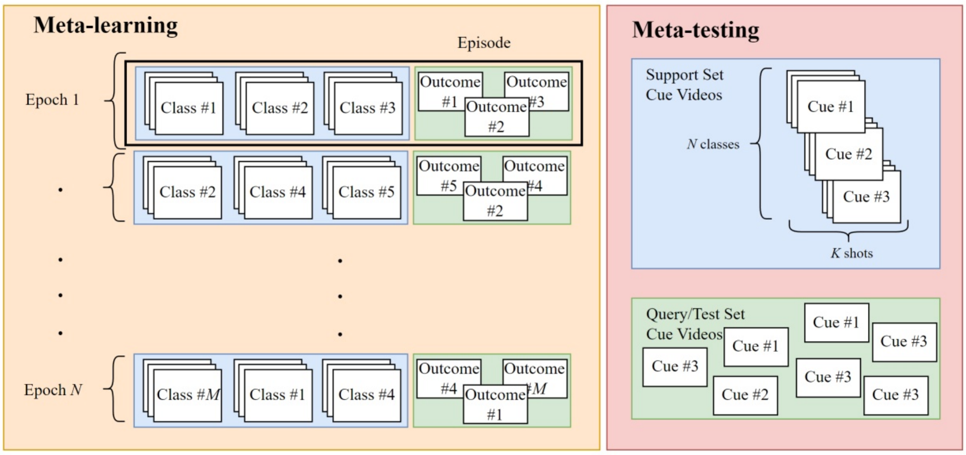


Figure 4.1: FSL meta-learning, where white boxes indicate video data. In the meta-learning phase, the model is trained by on a subset of the UCF101 dataset [Soomro et al., 2012], where a similar N classes and K shots are used as in the meta-testing phase. In meta-testing, the model is trained a final time on the support set before being tested on the test set.

4.2.2 Hyperparameters

As the underlying models in all experiments is the same, it is expected that hyperparameters are similar due to their shared foundation. To briefly reiterate, the loss function used was sparse categorical cross-entropy, and class weighting is applied to prevent the model from simply learning the distribution of the data.

When choosing the optimal optimizer for this experiment, Zhou et al. [2021] noted that Adam-type optimizers suffer from worse generalization performance than stochastic gradient descent. They uncovered that SGD is more locally unstable at sharp minima, and can better escape from them compared to flatter minima. Since meta-learning tasks require quick adaptations, an SGD optimizer is more suited in this experiment compared to the Adam optimizer used in the baseline experiment. However, from testing during the genetic search it arose that SGD did not perform as well as Adam during the meta-learning. Therefore, in the end the Adam optimizer was used against the interpretation of Zhou et al. [2021] due to its poor validation performance.

The parameters determined by the genetic search were: dropout at 0.5, 24 frames per snippet, batch size of 8, learning rate at 1e-2, no kernel regularization, freezing the backbone, and disabling causal convolutions. This time, the relatively high learning rate in the meta-testing phase stands out. However, the idea behind meta-learning was that the meta-learning prepares the model for quick convergence, which fits a higher learning rate. The meta-testing phase consists of 5 epochs.

The following hyperparameters, specifically related to the FSL approach, are the number of meta-tasks and the number of training shots per class. These parameters were also included in the genetic algorithm search. First, a high number of meta-tasks provides the model with more diverse range of learning iterations and therefore more frequent adaptations of its weights. This may benefit generalization to the target domain. The choice of the number of meta-tasks depends on timely termination of training before weights converge to a path that is not suitable for the target domain. In this experiment, 10 meta-tasks were selected by the genetic search from a possible 25. Each meta-task is trained for one epoch only, with a lower learning rate of 1e-4 to prevent early convergence.

Next, the number of shots indicates the number of examples per class in the meta training sets. This is important in preventing overfitting. The more shots are included, the more information the model receives to distinguish between the classes. The genetic search result in a 5-shot setting. However, during model construction, it was obvious that a 5-shot setting would lead to inferior results due to limited available samples. Therefore it has been upgraded to 10 shots in this experiment, which allows the model to learn efficiently without overfitting or training too rigorously on the meta-learning datasets. The 10-shot setting is used in both the meta-learning and meta-testing phase to increase relatedness between the two phases. A larger number of shots may not necessarily be beneficial due to the impact of the meta-learning phase on the weights convergence, so a balance must be struck between meta-tasks and shots.

4.2.3 Meta Training Tasks & Data

The meta-learning tasks are based on the UCF101 dataset (Section 3.1.8). This dataset was chosen due to its diverse range of video classes, providing a diverse and varied range of meta-tasks.

For each meta-task, $N = 24$ classes are randomly sampled from the UCF101 dataset. N is consistent with the number of classes annotated in the SLAPI dataset. For each of these classes, $K = 10$ random shots are selected. This process results in a diverse training set for each meta-task. In addition to this training set, a validation set is created for each task consisting of Q query images, using all available instances per class in the

UCF101 dataset. While performance on the validation set is not the primary focus of these experiments, they were monitored in hyperparameter selection to guarantee the meta-tasks had a positive impact on the model training. The data in the training and validation splits followed the guidelines by Soomro et al. [2012] to ensure a fair split that does not inflate performance due to poor sampling.

4.3 Active Learning Experiment

This section presents the final experiment, where the baseline approach is expanded upon using AL. The goal of these experiments is to investigate the improvements in cue classification that can be achieved when the model selects more informative snippets from the unlabeled pool using an intelligent sampling strategy, thereby keeping the labeling costs low. The outcomes from this experimental run will be compared to the results from the baseline and FSL experiments to provide an understanding of the potential advantages and limitations of this method.

4.3.1 Framework

This approach expands upon the standard motion analysis framework laid out in Section 4.1.1. The framework is illustrated in Figure 4.2. In this approach, the MoViNet A2 model is trained on the available instances in the labeled pool, allowing it to make predictions on the unlabeled instances using the feature representations built by its penultimate layer. Following some sampling strategy (discussed in Section 4.3.3), the unlabeled instances are selected for labeling by the oracle. The oracle then produces labeled samples, which must be removed from the unlabeled pool and appended to the labeled pool. This dataset is then shuffled and the MoViNet model is trained again. This process continues until convergence or time constraints in the user study are reached.

The core concept is that the model can provide samples more intelligently to the oracle after every iteration, enabling better training with use of only limited samples. After the iterative process is completed, the model is trained a final time on the completed dataset, resuming from where it left off in the previous iteration. It is then evaluated based on the metrics accuracy, precision, recall and F_1 score based on the test set.

4.3.2 Hyperparameters

Other than the hyperparameters specific to AL, the hyperparameters that are shared with the baseline MoViNet A2 model are outlined briefly. The loss function used was sparse categorical cross-entropy, the optimizer selected was Adam and class weighting was applied. Every new iteration, the class weights were updated based on the current labeled pool.

As for hyperparameters specific to this approach, the number of active learning iterations, the number of samples to be annotated per iteration were considered. First, the number of active learning loops regulates the iterations in which the unlabeled pool is sampled. It also functions as an upper limit as to how much the model can learn from the unlabeled pool. In this experiment, the number of iterations was limited to 3, and the number of instances that were sampled from the unlabeled pool was set to 50. This way, a

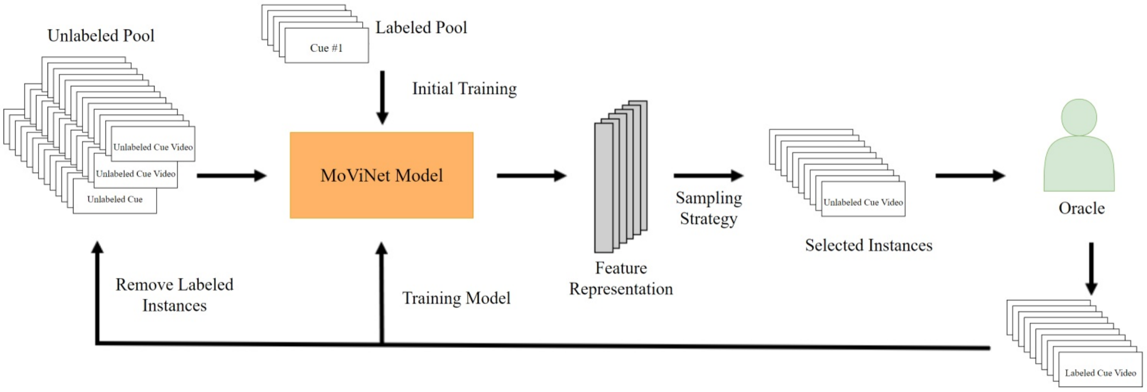


Figure 4.2: Illustration of the iterative process involved in AL, from initial training on the labeled pool, through the sampling strategy and labeling of new instances by the oracle, to retraining the model on the new labeled pool. White boxes indicate cue snippets.

batch of 50 samples were incorporated in the training dataset each iteration, allowing the model to learn using these new samples and more effectively select the subsequent batch of unlabeled samples. Selecting too few samples might slow the learning process, as the limited new information can hinder the model's ability to generalize. In contrast, choosing too many samples can introduce noise, adding an overload of new variations that might shift optimal parameters beyond the model's current training state. In total, 150 additional samples were labeled using active learning striking a balance with the preexisting labeled pool. These parameter settings were chosen as only sufficient data was available for roughly 150 samples.

To simulate the user study in the genetic search algorithm, the test set was used as unlabeled pool. The labels of the test set are known, and can be retrieved automatically in the active learning iterations. This resulted in the following hyperparameters: dropout at 0.3, 14 frames per snippet, batch size of 8, learning rate at 1e-3, no kernel regularization, freezing the backbone, and disabling causal convolutions. Other than the batch size, these parameters are identical to the baseline experiment. Given that the dataset was identical to the one used in the baseline experiment, the results were as expected. However, due to the training being distributed over four iterations of three epochs each, there was potential for discrepancies.

Lastly, the number of epochs in this setup is important, as it must be controlled in relation to the number of AL loops. Since the model is trained on the labeled dataset in each loop, allowing training for 5 epochs each iteration in the setup could lead to early convergence, neglecting potentially informative unlabeled samples. To prevent this, the number of epochs is set to a conservative value of 3 each iteration. After the active learning process is completed, the model is trained for another 3 epochs to consolidate the final samples.

4.3.3 Sampling Strategy

In this study, a pooling-based sampling strategy is used, as all the unlabeled data is available and can be evaluated for uncertainty measures based on which a predetermined number of instances is fed to the oracle for labeling. As was laid out in Section 2.3.2, multiple sampling strategies for selecting instances to feed the oracle were identified as reliable choices: uncertainty-based, entropy-based and best-versus-second-best. Infant behavior is often ambiguous, with multiple behaviors occurring simultaneously. As a result, it is beneficial to include instances in the model training that share these ambiguous properties.

However, the SLAPI dataset is largely unbalanced, which may lead to oversampling of the majority classes when using these approaches. To counter this, a stratified class-based approach is preferred. Instead of focusing on the predicted class of the unlabeled snippets, this method iterates over the classes and selects samples with the highest probability of belonging to that class. This allows for the localization of the underrepresented classes, even if they are not the predicted class. Algorithm 2 (Appendix A) outlines this sampling algorithm step-by-step. Duplicates may arise as samples could have the highest probability for multiple classes, which is prevented by only returning unique samples. However, this approach is not standard in related works, which commonly opt for standard uncertainty, diversity or best-versus-second-best sampling [Joshi et al., 2009]. Therefore, it is an interesting ablation condition to determine the impact of this problem-specific sampling strategy, compared to standard uncertainty sampling.

4.3.4 Snippet Generation

The unlabeled pool in the SLAPI dataset consists of lengthy, untrimmed videos, with a duration of up to hours. During manual annotation it was observed, as is detailed in Table 3.1, that typically behaviors of interest last from 2 to 10 seconds. Irrelevant behaviors and other camera intrusions can last from a few seconds to up to minutes, as they are not as intricate as the hunger and discomfort behaviors. By dividing the untrimmed videos into snippets of roughly 10 seconds, it is thought that the behaviors will be fully captured without making a snippet excessively long such that it captures multiple behaviors. Since preterm infants can show different cues simultaneously it is not possible to prevent multiple cues from occurring within the same snippet.

During AL, the goal is to find temporal boundaries of action instances and predict their labels based on such a snippet. To this end, Vahdani and Tian [2021] states that proposal generation is required to find the intervals at which cues occur. The proposal generator uses the MoViNet model of this experiment, and for each AL loop is initialized using the model that was enhanced from training. It walks through the untrimmed video frames at the FPS preferred by the model, which is 5 FPS in this case. After going through 3 seconds of video, a label is predicted for the snippet. The next 3-second snippet is processed. If the predicted label is the same as the previous snippet, it is assumed that they represent the same behavior and are grouped. This also means that snippets cannot overlap. Cues may not exactly fit the temporal boundaries this approach generates. However, this may be adjusted by the oracle when they are presented with the snippet. This process is halted after the combined snippet has a length of 15 seconds, since it was assumed that relevant behaviors take up to 10 seconds. This means that either the behavior was repeated or the classification was incorrect. In either case, a new snippet was created. The number of frames of the created snippet was reduced so that it is consistent with the labeled pool, retaining evenly spaced frames only. This algorithm returns the temporal coordinates indicated by frame starting and stopping index, and the processed snippet. Algorithm 3 (Appendix A) shows this proposal generator algorithm step-by-step.

4.3.5 User Study

This study used an oracle to provide labels for video snippets selected by the sampling strategy, while also allowing to oracle to modify snippets if needed. This could be required as the proposal generation algorithm uses a fixed 3-second interval, which may not fit the cues. Through command line input, the oracle could label snippets, skip snippets, modify snippets and replay snippets. The option to label snippets outside of the available classes was also presented, with the videos being stored outside of the labeled pool. To maximize the efficiency and minimize the redundancy in the annotation effort, the oracle's annotations were saved as snippets for future model training, effectively producing the dataset as byproduct of the study.

The oracle for the user study was a research intern employed at the UMC Utrecht. They were familiar with preterm infant hunger and discomfort cues, and agreed to participate in the study. Their task only involved providing the labels for snippets, and modifying snippet length whenever they deemed necessary to capture the entire cue.

Given the hyperparameters, the user study was divided into 3 iterations in which 50 samples were labeled. These samples were extracted from the pool of unlabeled and untrimmed videos, using the proposal generation algorithm. The proposal generator was reinitialized at the beginning of each iteration with the model that was retrained on the updated labeled pool.

The user interface of the study was implemented using OpenCV. The command line interface was native to the HPC environment the experiment was ran in. For each sample, the oracle was first shown the video. This showed the progress in seconds. The class labels and their indices were then printed, including instructions for how to skip a snippet, modify a snippet's duration, replay a snippet or give it a label outside the available classes. If the oracle entered incorrect input, they would be shown an error and requested to select a label within the available range. Figure A.1 in Appendix A shows the relevant command line and video interfaces, including instructions.

4.4 Ablation Studies

In this section, the ablation studies conducted to understand the impact each model component are set up. The modifications to the model can be divided into three categories: dataset, core component and AL modifications. Dataset modifications are only tested on the baseline main settings. The modifications to model components are also threefold. It involves regularization, training the model backbone and the application of causal convolutions. Although these techniques are either standard in machine learning, or MoViNet specific, they were not selected by the genetic search. A final modification explored is the importance of the sampling strategy in the AL user study. This aims to determine how stratified sampling impacts the model quality, compared to more conventional uncertainty sampling.

This means, in total, 12 model variations were included in the ablation study. Table 4.2 shows all possible combinations of component settings. Based on the evaluation metrics used throughout this study (accuracy, precision, recall and F_1) V0-V3 and V12 will be interpreted individually. V4-V10 will be compared with each other, possibly revealing synergistic effects.

Version	Backbone	Causal	Reg.	Model	Description
V0	—	—	—	Baseline	Increasingly larger dataset size.
V1	—	—	—	Baseline	10 data initializations/splits.
V2	✓	—	—	Baseline	Optical flow preprocessing.
V3	—	—	—	Baseline	Leave-one-out cross validation.
V4	✓	✓	L2	Baseline, FSL	—
V5	—	✓	L2	Baseline, FSL	—
V6	✓	—	L2	Baseline, FSL	—
V7	—	—	L2	Baseline, FSL	—
V8	✓	✓	—	Baseline, FSL	—
V9	—	✓	—	Baseline, FSL	—
V10	✓	—	—	Baseline, FSL	—
V12	—	—	—	AL	Unstratified uncertainty sampling

Table 4.2: Ablation study versions. Different combinations of component settings. Data modifications are applied to the baseline main settings, while model modifications are compared against each other with all possible options. Models V0-V3 use the main settings of the baseline approach. Models V4-V11 are tested on both the baseline and FSL approaches, deviating from their main settings. Model V12 is tested on the AL main settings. V11 is omitted as its settings are identical to the main settings.

4.4.1 Data Modifications

The first adjustment involves systematically increasing the number of training shots per class (V0). Starting from 1, the shots are increased to 2, 3, 5, 10, 15 and 25. When the number of shots is larger than the available instances in that class, all available instances are used. This does make it more difficult interpret the performance based on more data, as at some point only few classes may gain new training samples. In machine learning, increasing the amount of data used and its performance can be described as a diminishing returns curve. Initially when the model only has access to a small set of data, adding more data yields larger performance increases than when the dataset is already sufficiently large. While it is common for medical datasets to lack availability and quality annotations, it is not yet determined what role this limitation has in this thesis. By increasing the size of the dataset and plotting the performance, it can be determined what the state of the current dataset is, and how increasing the number of instances may improve performance.

A further component is the data initialization (V1). Since the data is not sufficiently uniform, the train, validation and test splits may impact performance. To get a clearer understanding of the impact of the data initialization, the split is done 10 time completely at random. Each time the model is trained using the main settings on this dataset. This way, the effect of a new random initialization on performance can be mapped.

In the experiments, preprocessing consisted of cropping and rotating the videos to maintain consistency and relevant information in the center of the video. Then, the frames were reduced to preferred resolution of the MoViNet A2 model and samples in steps of 5 FPS, also preferred by the model. In this ablation study (V2), optical flow in combination with the 1-dimensional gray-scale frame, resulting in a 3-dimensional representation. Optical flow may be beneficial to the model, since it captures motion information, which is important to cue interpretation. Since MoViNet models are pretrained using RGB data, the model backbone was set to trainable, allowing for fine-tuned feature representations.

Multiple popular dense optical flow approaches exist, with Horn-Schunck [Horn and Schunck, 1981] and Farnebačk [Farnebäck, 2003] being the most common. Both methods compute the optical flow for each pixel in the input frame. However, due the polynomial expansion approach in Farnebačk's method, the optical flow is more robust to noise, large displacements and illumination changes. This does mean the method is more computationally expensive. Due to the noisy nature of the SLAPI dataset, Farnebačk was selected as optical flow method. Farnebačk's algorithm returns horizontal and vertical displacement values for each pixel compared to the previous frame. These values were not transformed into polar coordinates due to the natural discontinuity in angle representation, where the range of angles extends up to 360 degrees. Although there is only one degree difference between 0 and 360 degrees, this is not adequately represented in the numerical value.

To determine the importance of the infant's behavior in the classification of their behaviors, a leave-one-out ablation was tested (V3). One infant's data was used as test set, while the other infants' data comprised the training and validation set. This approach provides insight into the generalizability of the model, but also of the cues across infants. Infant 65 was left out of this ablation, due to its limited exhibited behaviors.

4.4.2 Model Component Modifications

By training the model backbone, not only the final classification layer is trained, but all layers of the model. This way, the weights of the backbone are fine-tuned, acting as domain adaptation of sorts. However, this introduces increased risk of overfitting due to the increased number of parameters that are now adapted based on the imperfect dataset.

The model's performance was evaluated with and without L2 regularization to understand the impact of regularization on performance. Regularization techniques help prevent overfitting by adding a penalty to the loss function. It was thought due to the complexity of the MoViNet model, this would be beneficial. However, it was not found to be best by the genetic search.

Causal convolutions, one of the hallmarks of the MoViNet architecture (see Section 3.3), were not enabled in the architecture found by the genetic search algorithm. Therefore, their impact on the models is tested in the ablation study. They are especially useful for applications with high time and memory complexity, such as motion analysis.

All these combinations were combined into 7 new model versions, and they were tested in both the baseline and FSL approaches.

4.4.3 Active Learning Modifications

Considering the importance of sampling strategies in addressing the underrepresented classes to arrive at a balanced model, stratified probability-based sampling was adopted. However, it is important to validate the effectiveness of this approach by comparing it to standard sampling strategies, like uncertainty-based sampling. This comparison will provide insight into how different strategies can impact model performance. Due to the requirement of a user study, only the uncertainty-based sampling strategy was tested for 3 iterations and 50 samples per iteration, similar to the main experiment.

5 Results

This section presents the results of the experimental runs. The results of the main experiments are summarised in Table 5.1, and described in the relevant subsections. The table also includes the metrics for random and weighted guessing classifiers, where the weighted classifier uses the distribution of the training set as prior probability. These metrics serve as an indication of task difficulty, and a point of comparison. No metric computed on the test set exceeds .10.

For each experiment, the metrics validation accuracy, accuracy, precision, recall and F_1 are computed on the test set. The training history of these metrics on the training and validation set is shown in Appendix B, as well as a confusion matrix of the predictions on the test set. The results of the ablation studies are reported in the final subsection.

Experiment	**Val. Acc.**		**Acc.**		**P**		**R**		F_1		**Trials**
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	
Baseline	.40	.02	.36	.02	.30	.03	.31	.03	.27	.03	10
FSL	.29	.03	.28	.03	.25	.03	.26	.03	.21	.02	10
AL	.38	—	.35	—	.28	—	.30	—	.26	—	1
Random guess	—	—	.04	.01	.04	.01	.04	.02	.04	.01	100
Weighted guess	—	—	.08	.02	.04	.01	.04	.01	.04	.01	100

Table 5.1: Summary of experimental results for different models. Each model's performance is assessed in terms of Validation Accuracy (Val. Acc.), Accuracy (Acc.), Precision (P), Recall (R), and F_1 score. For each metric, the Mean (M) and Standard Deviation (SD) across 10 trials are provided, where there was only one trial for AL.

5.1 Baseline Experiment

In the baseline experiment, the model underwent a training phase spanning 5 epochs. Upon completion, it achieves a mean validation accuracy of .40(SD=.02) across ten 10 trials. The model's training history, along with other validation metrics are outlined in Appendix B, Figure B.1. When evaluated on the test set, a reduction in the model's mean accuracy is observed, dropping to .36(SD=.02). The mean precision is found to be .30(SD=.03), demonstrating that an average of 30% all predicted positive instances were accurately classified. The mean recall rate is measured at .31(SD=.03), indicating that the model correctly detected 31% of all positive examples in the test set, averaged per class. The mean F_1 score is calculated to be .27(SD=.03). The relatively small standard deviations for all metrics suggest that the 10 executed experimental trials yield consistent results, summarised in Table 5.1.

A more nuanced understanding can be gathered by examining cue-level scores, as opposed to solely focusing on overall model performance. As illustrated by the confusion matrix in Figure 5.1, performance varies with behavioral cues. While the results from this confusion matrix are drawn from a single trial of the baseline model experiment, which requires cautious interpretation, it offers an insightful visual representation that helps in understanding the model's performance. Among the various cues, "*Adult hand*", "*Bottle feeding*", "*Bottle sucking*", "*Crying*", "*Hand sucking*", "*Shiver*", "*Still*" and "*Tugging*" are all relatively accurately predicted with 50% recall or higher. The cues "*Bottle feeding*",

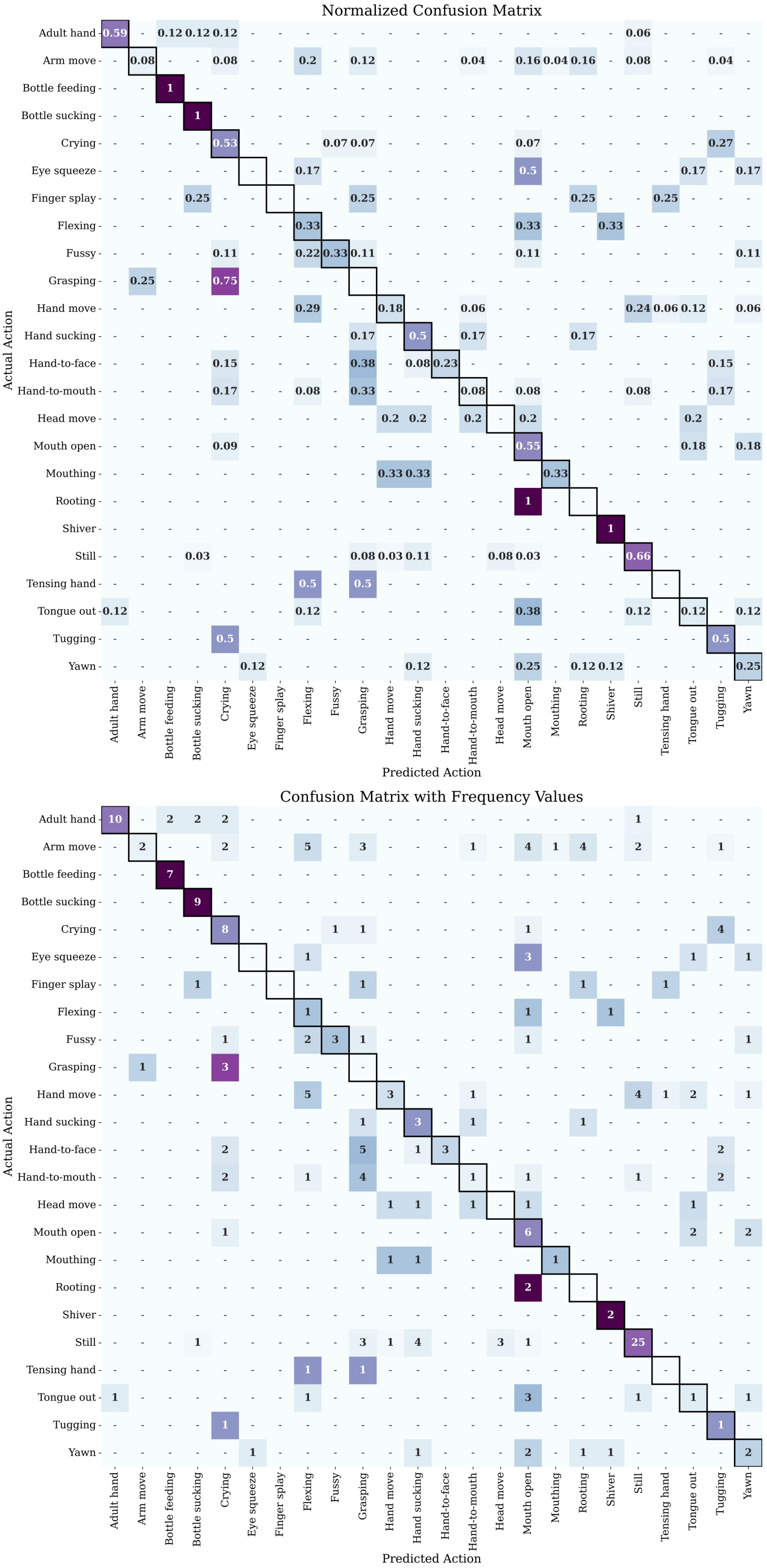


Figure 5.1: Confusion matrices for baseline experiment first run with MoViNet A2. On top, a normalized confusion matrix displaying the proportion of correct and incorrect predictions for each class. On the bottom, a frequency-based confusion matrix showing the absolute number of predictions for each class.

Class	**P**		**R**		F_1		**Samples**
	M	*SD*	*M*	*SD*	*M*	*SD*	
Adult hand	.89	.08	.77	.07	.82	.05	17
Arm move	.24	.21	.08	.08	.11	.09	25
Bottle feeding	.96	.07	.89	.09	.92	.04	7
Bottle sucking	.74	.07	.99	.03	.84	.05	9
Crying	.36	.05	.51	.10	.42	.05	15
Eye squeeze	.10	.11	.17	.21	.11	.10	6
Finger splay	.00	.01	.03	.08	.01	.02	4
Flexing	.04	.05	.18	.23	.06	.07	3
Fussy	.61	.26	.23	.15	.29	.12	9
Grasping	.00	.00	.00	.00	.00	.00	4
Hand move	.42	.18	.26	.18	.30	.17	17
Hand sucking	.20	.05	.57	.14	.29	.05	6
Hand-to-face	.39	.36	.21	.15	.23	.14	13
Hand-to-mouth	.29	.13	.21	.12	.21	.08	12
Head move	.00	.00	.00	.00	.00	.00	5
Mouth open	.13	.10	.18	.19	.14	.13	11
Mouthing	.28	.30	.26	.14	.24	.16	3
Rooting	.00	.00	.00	.00	.00	.00	2
Shiver	.52	.34	.67	.41	.55	.32	2
Still	.77	.06	.55	.07	.64	.05	38
Tensing hand	.01	.03	.11	.21	.03	.05	2
Tongue out	.14	.12	.12	.10	.13	.10	8
Tugging	.05	.05	.28	.34	.08	.09	2
Yawn	.16	.14	.14	.11	.15	.12	8
Hunger	.37	—	.37	—	.33	—	80
Discomfort	.10	—	.20	—	.13	—	30
Other	.34	—	.31	—	.32	—	118

Table 5.2: Performance metrics for individual classes in the baseline experiment, displayed in terms of Precision (P), Recall (R), and F_1 score. For each metric, Mean (*M*) and Standard Deviation (*SD*) values are provided. The number of samples per class is also included to illustrate data distribution. This class-based evaluation allows for a more nuanced interpretation of the model's performance.

"*Bottle sucking*" and "*Shiver*" achieve notably high recall scores of 100%. Interestingly, classes that demonstrate high recall scores in the confusion matrix include both majority and underrepresented classes. For instance, "*Still*" consists of 149 samples. "*Bottle sucking*", on the other hand, consists only 31 samples. Its 9 test samples are all correctly identified. This suggests that the impact of class imbalance is mitigated to an extent. Further evidence supporting this claim is provided by the absence of preferred classes for prediction.

Table 5.2 provides a more comprehensive representation of the classification performance across different classes. The metrics precision, recall and F_1 are reported for each class. Additionally, the test set sample size is reported for each class, to illustrate the distribution of the data. The standard deviations are reported to shed light on how sensitive classes were to different model runs.

When considering the class "*Adult hand*" (P=.89, R=.77 and F_1=.82), the model returns high scores in all evaluation metrics. The high precision score suggests that only few false positives were predicted, whereas the recall score indicates the model correctly identifies 77% of all snippets where an adult hand appears in the frames. Due to the nature of this cue, in which a foreign object enters the frame, it is unsurprising that the model performs well. The cue can be classified based not only on the intricacies of the infant's behavior, but also the presence of other objects. Similar cues, such as "*Bottle feeding*" (P=.96, R=.89 and F_1=.92) and "*Bottle sucking*" (P=.74, R=.99 and F_1=.84) also demonstrate good performance, which can likely be attributed to the presence of a foreign object.

Conversely, "*Arm move*" (P=.24, R=.08 and F_1=.11) demonstrates markedly low precision, recall, and F_1 scores, implying that the model struggles to identify this class correctly and differentiate it from other cues. For instance, confusions occur where "*Arm move*" instances are predicted as being "*Flexing*", "*Mouth open*" or "*Rooting*". An instance of "*Grasping*" is incorrectly classified as "*Arm move*". Next, "*Finger splay*" (P=.00, R=.03 and F_1=.01) scores hardly any precision and very low recall and F_1 scores, which suggests the model is not able to identify these classes correctly. Its instances are incorrectly classified as being "*Grasping*", "*Rooting*" or "*Tensing hand*" This also appears to be the case for "*Flexing*" (P=.04, R=.18 and F_1=.06), "*Tensing hand*" (P=.01, R=.11 and F_1=.03) and "*Tugging*" (P=.05, R=.28 and F_1=.08). Poor performance across all metrics signifies that the model frequently misclassifies or fails to detect these classes.

The classes "*Grasping*", "*Head move*" and "*Rooting*" all stand out as the model appears unable to correctly identify this class at all (P=.00, R=.00 and F_1=.00), reflected in its zero precision, recall, and F_1 scores. The limited number of test samples (4, 5, and 2 respectively) for these classes means that there is only limited training data available. This might have contributed to the poor performance, suggesting a need for a more balanced dataset to properly interpret the results for these cues. An alternative explanation is the ambiguous nature of the classes. For example, grasping involves the movement of the arm and hand, for which separate classes exist. Similarly, rooting involves movement of the head, moving the hands towards the mouth and displaying sucking-like behaviors. This creates challenging circumstances for the model to distinguish between these classes. This is particularly true for "*Arm move*" as it has a relatively high number of samples, and could have usurped smaller classes explaining poor performance in the metrics of both classes.

The model also struggles with more ambiguous or subtle classes such as "*Eye squeeze*" (P=.10, R=.17 and F_1=.11), "*Fussy*" (P=.61, R=.23 and F_1=.29) and "*Head move*" (P=.00, R=.00 and F_1=.00). These relatively low scores can be explained by the inherent ambiguity or complexity of these behaviors. Fussiness is defined as rapid movements of the head, arms and changes in body tension. So the class consists of many different behaviors grouped together, which somewhat overlap with other classes as well. "*Eye squeeze*" is a subtle tensing of the eyelids, which is often accompanied by an open mouth and flexing of the hands, just like "*Head move*" is often combined with movements of the eyes and arms. These classes also have relatively few samples, suggesting that their complexities cannot be learned from the available data.

Finally, the class with the most samples is "*Still*" (P=.77, R=.55 and F_1=.64). Its high precision and recall scores are useful to ensure that when an infant is genuinely still, and requires no care, it is accurately recognised. The scores observed for this class do follow that pattern, which is a promising outcome.

When examining the standard deviations, some issues become apparent in the following classes, with respect to precision: "*Hand-to-face*" (M=.39, SD=.36) and "*Mouthing*" (M=.28, SD=.30). The considerable inconsistencies in precision suggests possible variability within the class samples. While the number of samples is sufficient, this variability could have been learned in some experimental trials, but not in others. Since these are behavioral cues potentially indicative of hunger, their detection is important. However, particularly when taking the low recall scores into account (M=.21 and M=.26 respectively), the classification of these classes is lacking in reliability. Similarly, substantial variability in recall presents challenges for the following classes: "*Shiver*" (M=.67, SD=.41) and "*Tugging*" (M=.28, SD=.34). The variable identification of these classes again indicates unreliable detection. This may be due to dataset limitations or the complexity in the behaviors. Nevertheless, since these behaviors are not indicative of hunger or discomfort, their detection is less vital. Furthermore, the precision and recall scores are relatively high, despite their large standard deviations.

To generate the previous results, a cue-based approach was used. Table 3.1 provides a tentative categorization into the categories of interest: "*Hunger*", "*Feeding discomfort*" and "*Other*". The scores for the metrics precision, recall and F_1 can be averaged to derive category-based scores. Using this categorization, "*Hunger*" scores P=.37, R=.37 and F_1=.33. "*Feeding discomfort*" scores P=.10, R=.20 and F_1=.13. Finally, "*Other*" P=.34, R=.31 and F_1=.32. This result is still problematic, as precision and recall are lower than .50. So when using the tentative categories, less than half of the cue snippets are successfully classified. Once again, this issue may be attributed to the lower availability of samples in those classes. Upon examining the confusion matrices in Figure 5.2, it is apparent that substantial confusions occur across the three categories. The model distributes it predictions almost evenly across the categories, arriving at a total accuracy of .55. This result is comparable to that of majority class voting for "*Other*", which would yield an accuracy of .52 (118/228), thereby suggesting that the baseline approach was not effective in broadly classifying hunger or discomfort behaviors. Instead, it is more productive focusing on individual cues, and using the per-class performance to guide the interpretation.

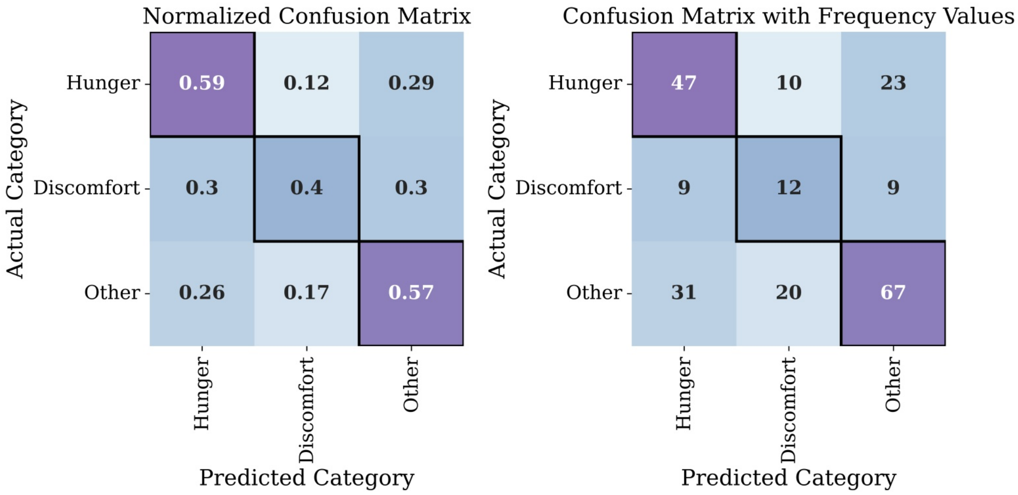
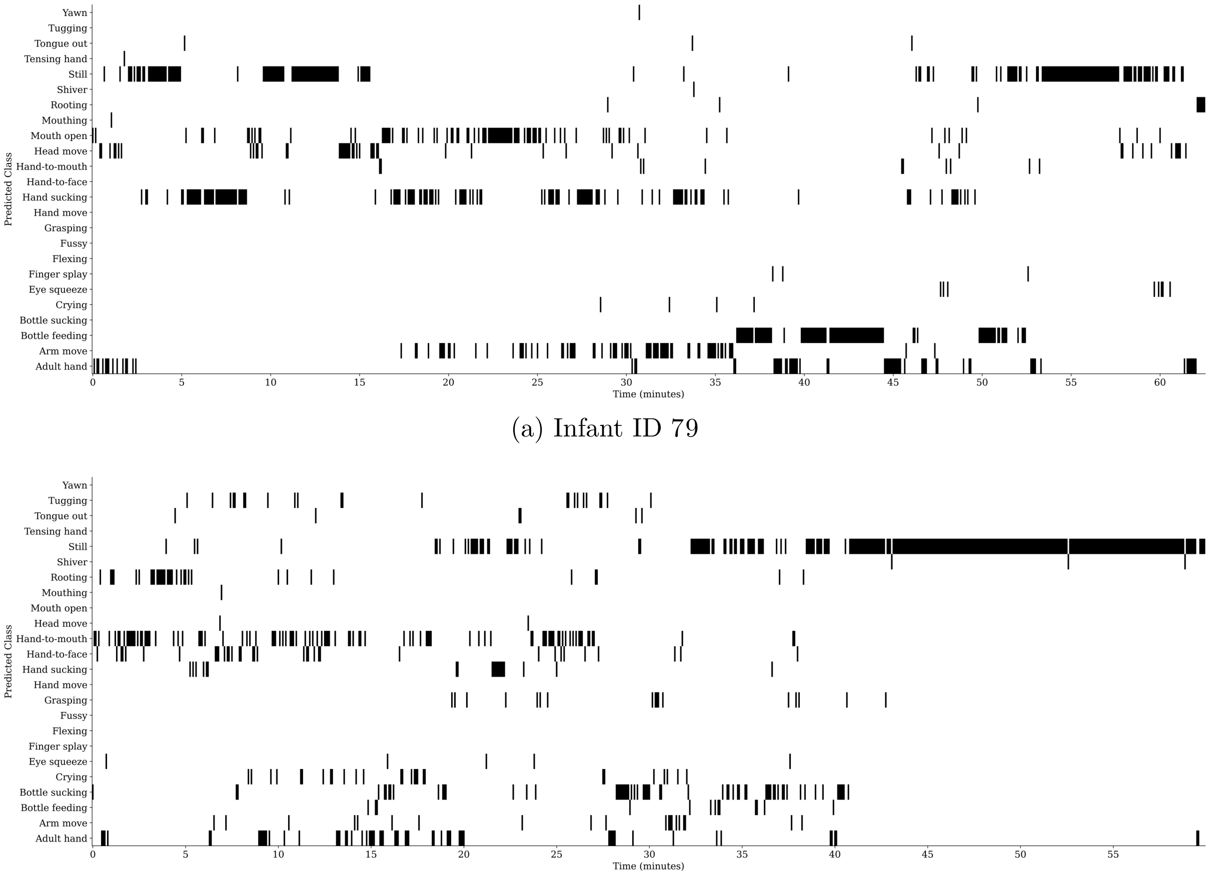


Figure 5.2: Confusion matrix for the baseline reduced to three main categories of behavior, being "*Hunger*", "*Feeding discomfort*" and "*Other*". This labeling results in .55 accuracy for the first run of the baseline experiment.



(a) Infant ID 79

(b) Infant ID 95

Figure 5.3: Predicted classes over time for infant 79 (a) and infant 95 (b). These plots showcase the behavior classifications of each infant across the duration of their respective videos. Despite low metric scores, they demonstrate the model's capability in consistently detecting significant behaviors such as "Bottle feed", "Bottle sucking", "Adult hand", and periods of stillness. The plots also indicate periods of restlessness in both infants, visible through the wide range of predicted behaviors.

To put these results into practice, the entire videos of infants 79 and 95 were fed to the model, with consecutive snippets of 4.8 seconds being classified. The predictions are shown in Figure 5.3(a) for infant 79 and Figure 5.3(b) for infant 95. Due to the low metric scores, the exact classifications are not the primary points of interest in these plots. Both infants 79 and 95 both exhibit a wide range of behaviors from the start of the video, irrespective of whether the classifications are inaccurate. This suggests the infant is restless, potentially due to hunger. After approximately the 37-minute mark, infant 79 is bottle-fed. It also appears as though the model successfully recognised the "*Adult hand*" and "*Bottle feeding*" in the video, as those classes are consistently predicted over a 15 minute stretch. After the feeding stops around the 45-minute mark, the infant continues to display signs of restlessness, as indicated by the wide range of predicted behaviors. However, after a second feeding after minute 50, the infant's actions are largely predicted as "*Still*". Any remaining deviations from "*Still*" could be due to movements whilst sleeping, or inconsistencies in the data. In a similar manner, the model successfully identifies the introduction of a pacifier to infant 95 at approximately the 28-minute mark in the video. At this same time, the infant was tube-fed. From the 41-minute mark onward, the infant appears to have fallen asleep, which is common after feeding. As a result, there is an apparent absence of predictions for any class other than "*Still*". Thus, the reliable performance in recognising "*Bottle feed*", "*Bottle sucking*", "*Adult hand*" and "*Still*" results in an accurate description of infants' untrimmed videos, despite the possible misclassifications of other behaviors. All these observations, drawn from the plots, were validated by manual inspection of the untrimmed videos.

5.2 Few-Shot Learning Experiment

In the FSL experiment, the model underwent training for 5 epochs after pretraining the parameters on 10 meta-tasks. This results in a mean validation accuracy of .29(*SD*=.03) after 10 trials. The model's training history, along with other validation metrics are outlined in Appendix B, Figure B.2. When evaluated on the test set, the model's mean accuracy remains consistent at .28(*SD*=.03), meaning the model was able to accurately classify 28% of all instances. Mean precision is measured at .25(*SD*=.03), while recall is measured at .26(*SD*=.03). This indicates that the predicted positive instances were 25% are correct, while 26% of all positive samples are detected successfully, averaged across the classes. Finally, the test F_1 score is .21(*SD*=.02). Again, the relatively small standard deviations demonstrate consistency across the model runs.

When comparing these results to the baseline experiment, it is observed that the FSL model reveals lower metrics across the board. In the validation accuracy, there is a .11 decrease (M=.40 versus M=.29), while performance is decreased by .08 for accuracy (M=.36 versus M=.28). For precision (M=.30 versus M=.25) and recall (M=.31 versus M=.26) performance is reduced by .05 and F_1 (M=.27 versus M=.21) shows a difference of .06. This suggests that the meta-learning phase only hampers performance.

In terms of class-based performance (see Figure B.4 in Appendix B), there is a similar pattern in the performance per class, where majority classes "*Adult hand*" and "*Still*" perform relatively well. Meanwhile, underrepresented classes such as "*Finger Splay*", "*Rooting*" and "*Mouthing*" remain with low metric scores. This pattern suggests that despite the inherent countermeasures to manage class imbalance and functioning with limited data, theFSL framework does not significantly improve the detection of underrepresented classes. This could be due to data limitations, or variations within classes, indicating that simply changing the machine learning approach is not sufficient for overcoming this limitation and further improving performance.

The core idea behind FSL is to find a path in the parameter space that allows for quick fine-tuning of these parameters to an unspecified target domain. From the training history (Figure B.2 in Appendix B), however, it is observed that the validation loss actually starts out higher around 3.0 and does not recover from this deficit despite an early drop after the first epoch. As a result, model performance on the test set disappoints. Furthermore, there is no indication that underrepresented classes are more accurately detected. The genetic search algorithm returned the best learning rate at 1e-2. This is a comparatively high value, possibly resulting in undoing earlier learned weight updates, and overshooting minima in the loss landscape.

5.3 Active Learning Experiment

In this experiment, the model underwent training for 3 epochs before AL, and then for 3 epochs after each of the iterations. This resulted in a validation accuracy score of .38. The model's training history, along with other validation metrics are outlined in Appendix B, Figure B.3. When evaluated on the test set, the model's accuracy is measured at .35, meaning the model was able to accurately classify 35% of all instances. Precision is measured at .28, while recall is measured at .30. This indicates that the predicted positive instances were 28% were correct, while 30% of all positive samples were detected successfully, averaged across the classes. Finally, the test F_1 score is .26.

To see how the model is affected by the AL iterations, 24 samples from the test set were set aside. One for each class. Figure 5.4 shows the evolution of the certainty over these samples over the course of the learning phase. The probabilities of each sample belonging to a class are shown after each iteration. It can be seen that initially before the first AL iteration, the probabilities are more spread out across the grid. This indicates that the model has a higher degree of uncertainty in evaluating these samples. After the first iteration, the uncertainty is somewhat reduced, as showcased by the lighter colors occurring. This effect becomes stronger in the second and third iteration, where in many cases the model only considers a few classes per instance. It is noticeable in the last heatmap that "*Arm move*" is considered likely for many instances, where "*Mouth open*" also scores high probabilities. It appears that for these randomly selected samples, the majority of predictions are be inaccurate, as the highest probabilities do not align with the diagonal. This effect is not corrected for across training. Nevertheless, this examination effectively shows the evolution of probabilities across the iterations, and how the distribution of probabilities becomes more sparse.

When compared to the other experiments, the baseline model delivers slightly higher performance across all metrics, with a validation accuracy of .40 (versus .38 of AL), an accuracy of .36 (versus .35), a precision of .30 (versus .28), a recall of .31 (versus .30), and an F_1 score of .27 (versus .26). This indicates that while the AL model exhibits slightly lower metrics, it remains competitive with the baseline. In contrast, the FSL model posts lower metrics across the board, with a validation accuracy of .29 (versus .38 of AL), an accuracy of .28 (versus .35), a precision of .25 (versus .28), a recall of .26 (versus 0.30), and an F_1 score of 0.21 (versus 0.26). This comparison emphasizes the AL model's superior performance relative to the FSL model. When looking at the confusion

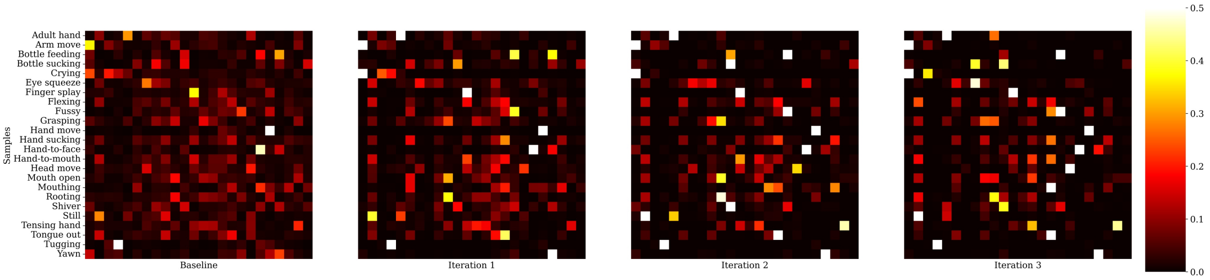


Figure 5.4: A series of heatmaps shows the evolving class probabilities for 24 samples across AL iterations using stratified maximum probability sampling. The y-axis orders samples by class, while the x-axis represents class probabilities. The heatmaps chronologically illustrate the progression of these probabilities.

matrices (Figure B.5), there are no stark differences between the approaches. All models have trouble with identifying classes like "*Flexing*", "*Eye squeeze*" and "*Grasping*". On the other hand, "*Adult hand*" and "*Still*" perform well once again.

To interpret how well AL has solved the issue with limited data, performance on the bottom 10 classes is examined. Table 5.3 compares the scores for the bottom 10 classes by number of training samples for the baseline and AL approach. The results of ablation version V12 are also included in this table. It can be seen that that performance on these classes remains disappointing. Only "*Mouthing*" (P=.11, R=.33 and F_1=.17) and "*Shiver*" (P=.25, R=1.0 and F_1=.40) are correctly predicted at least once. However,

Class	**Baseline**			AL			**Ablation V12**			**Samples**
	P	R	F_1	P	R	F_1	P	R	F_1	
Eye squeeze	.10	.17	.11	.00	.00	.00	.00	.00	.00	19
Finger splay	.00	.03	.01	.00	.00	.00	.00	.00	.00	11
Flexing	.04	.18	.06	.00	.00	.00	.00	.00	.00	10
Grasping	.00	.00	.00	.00	.00	.00	.00	.00	.00	11
Head move	.00	.00	.00	.00	.00	.00	.00	.00	.00	17
Mouthing	.28	.26	.24	.11	.33	.17	.09	.33	.14	12
Rooting	.00	.00	.00	.00	.00	.00	.00	.00	.00	8
Shiver	.52	.67	.55	.25	1.0	.40	.18	1.0	.31	8
Tensing hand	.01	.11	.03	.00	.00	.00	.00	.00	.00	5
Tugging	.05	.28	.08	.00	.00	.00	.17	1.0	.29	8

Table 5.3: Comparison precision, recall, and F_1 scores for the least represented 10 classes, using AL stratified sampling and ablation V12 uncertainty sampling. The data demonstrates that the stratified sampling method does not provide any apparent benefit in learning from underrepresented classes.

Class	**Iteration 1**	**Iteration 2**	**Iteration 3**	**Total**
Arm move	6	3	11	20
Crying	—	—	3	3
Eye squeeze	—	1	4	5
Finger splay	1	—	—	1
Fussy	2	—	6	7
Grasping	—	1	1	2
Hand move	5	1	4	10
Hand-to-face	1	—	3	4
Hand-to-mouth	2	—	2	4
Head move	1	—	5	6
Shiver	—	2	1	3
Rooting	—	—	3	3
Still	32	42	8	82

Table 5.4: The distribution of classes identified during each iteration of the AL experiment. The table highlights the frequency of each class per iteration and provides a total sum of instances across all iterations. The class "*Still*" was the most frequently identified across the iterations, with a total of 82 instances, while "*Arm move*" and "*Hand move*" were the next most common, with 20 and 10 instances respectively.

it was expected that the sampling strategy would shine more light on these classes to improve performance. When examining what labels the sampling strategy managed to retrieve, shown in Table 5.4, it turns out that some minority classes were retrieved from the unlabeled pool. Classes like "*Rooting*" (3 samples), "*Eye squeeze*" (5) and "*Head move*" (6) are labeled more than three times, but did not result in correct predictions. It is most likely that unlabeled pool follows the distribution of the labeled data, where classes like "*Still*" (82) and "*Arm move*" (20) are also more common. Therefore they also end up being added to the dataset most, and training may not be affected by the minimal influx for the underrepresented classes. When comparing these results against the baseline approach, a similar pattern arises. It appears that across the 10 trials some cues do get correctly identified. This is evidenced by the nonzero values in all classes but "*Grasping*", "*Head move*" and "*Rooting*". The class "*Shiver*" also stands out due to its comparatively high precision, indicating that the AL approach may have reduced performance in this class by including the additional samples.

The fact that performance is reduced in some classes may be caused by the inclusion of new samples from a different source video. It was previously established that the source video has an impact on generalization. To test this idea, and to gain insight in the quality of the samples, the samples gathered in the AL study are used as training set, while the original training set is now used as test set. To facilitate this approach, classes which were not sampled in the AL iterations are omitted. The baseline model is used for this experiment. Table 5.5 shows the class-based performance. It can be observed that performance is generally worse compared to the results class-based performance shown in Table 5.2. No correct predictions are made for "*Eye squeeze*", "*Finger splay*", "*Grasping*" and "*Hand-to-face*". It appears "*Still*" is the only class that outperformed the baseline, with a precision score of .93. All other classes achieve similar but significantly lower scores across all metrics compared to the baseline approach. This suggests that the samples generated from the AL iterations do not provide high quality training information from which the model can improve, as they do not generalize well to the existing dataset. It must be kept in mind that this classification task is less complex with only 13 classes, which may have inflated performance compared to the more complex 24-class problem.

Class	**P**	**R**	F_1
Arm move	.20	.04	.07
Crying	.08	.22	.12
Eye squeeze	.00	.00	.00
Finger splay	.00	.00	.00
Fussy	.10	.28	.15
Grasping	.00	.00	.00
Hand move	.20	.10	.13
Hand-to-face	.00	.00	.00
Hand-to-mouth	.18	.17	.17
Head move	.06	.12	.08
Rooting	.03	.25	.06
Shiver	.00	.00	.00
Still	.93	.36	.52

Table 5.5: Performance metrics for different classes using active learning samples as the training set and the original training set as the test set, including precision (P), recall (R), and F_1 score (F_1).

5.4 Ablation Studies

In this section, the results of the ablation studies are reported. The aim was to shed light on the model's behavior under various conditions and identify the key drivers of its performance. This rigorous evaluation helps understand the robustness of the model and guide future developments and improvements. The details of every ablation condition can be found in Table 4.2.

5.4.1 Results Data Modifications

Under ablation condition V0, the number of maximum training samples was gradually increased to examine the curve of diminishing returns and gain insight into how more data would have improved performance. This was tested using the baseline model. The plotted results in Figure 5.5 clearly show a positive trend of improvement in the evaluations metrics when increasing the maximum number of samples per class. After using all available data, and the AL gathered data, performance decreases. Across all metrics, performance is least favourable when only using a single training sample per class. There are notable leaps in performance when allowing 2 and subsequently 3 training samples per class. Interestingly, recall suffers a slight decrease, when using 5 samples. However, it recovers when increasing the number of samples further. Other than a minor reduction going from 10 to 15 samples, it is premature to conclude that this is due to diminishing returns as performance increases again when using the maximum of 25 classes per samples. When using the entire dataset (samples = 417, min. class = 4, max. class = 74), a drop in performance is observed. This drop in performance does not meet the expectations of diminishing returns, where positive returns are still the norm. It appears as though lifting the limit of training samples allows the imbalance in training data to become too large, and hindering training in the process. A mild recuperation in performance is observed when the data from the AL study was integrated in the dataset. While there are signs that the rate of improvement in some metrics is decreasing with additional data, it's not clear-cut across all metrics, and an improvement is still observed. As such, it would be beneficial to continue exploring ways to increase the dataset size, diversify the data, or improve the model architecture and training process to improve performance.

Ablation condition V1 was conducted to examine the impact of various data splits on the performance of the baseline model. Due to the limited dataset size, it was expected that the model performance could be affected by which samples end up in the training set. The use of 10 different data initializations results in slight but consistent improvements across all metrics, when compared to the baseline model. The mean validation accuracy is similar for the random initialization (M=.39, SD=.03) and baseline approach (M=.40, SD=.02), while V1 (M=.38, SD=.02) outperforms the baseline approach (M=.36, SD=.02) in mean accuracy on the test set. For precision (M=.30, SD=.02 versus M=.30, SD=.03), recall (M=.32, SD=.02 versus M=.31, SD=.02) and F_1 (M=.28, SD=.02 versus M=.27, SD=.03) the ablation study outperformed the baseline experiment. While these differences are only minimal, they suggest that the main initialization is not the most optimal split in the dataset.

For ablation condition V2, the preprocessing of the data was adapted. Rather than using the three RGB channels of the snippets, optical flow was computed. This resulted in 3-channel representations. Two channels were used for the horizontal and vertical optical

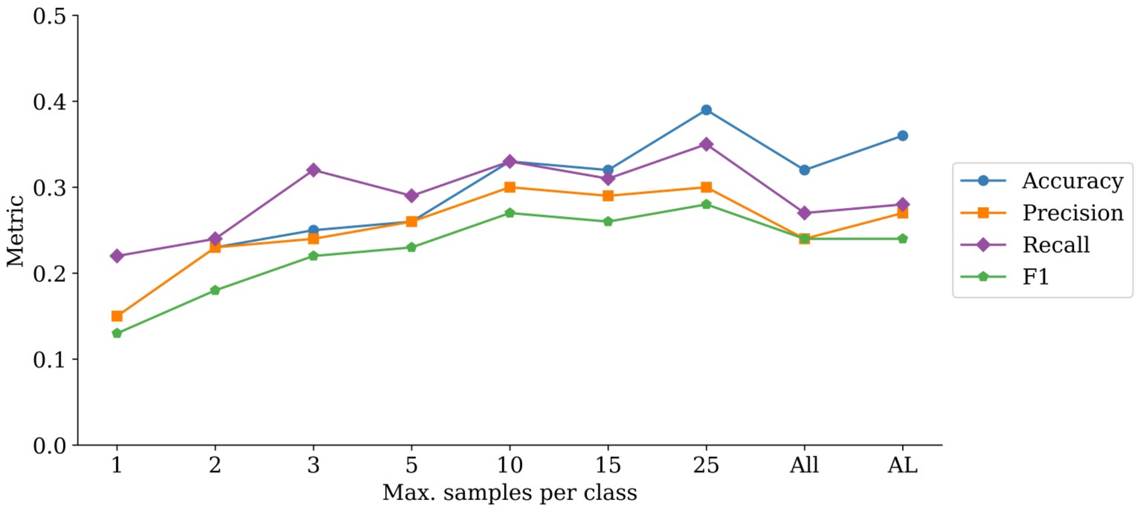


Figure 5.5: V0 — The plot illustrates the progression of metrics accuracy, precision, recall, and $F1$ in response to the increasing maximum number of samples per class. The figure clearly shows a general trend of improvement in these performance metrics with the growth in the number of samples per class.

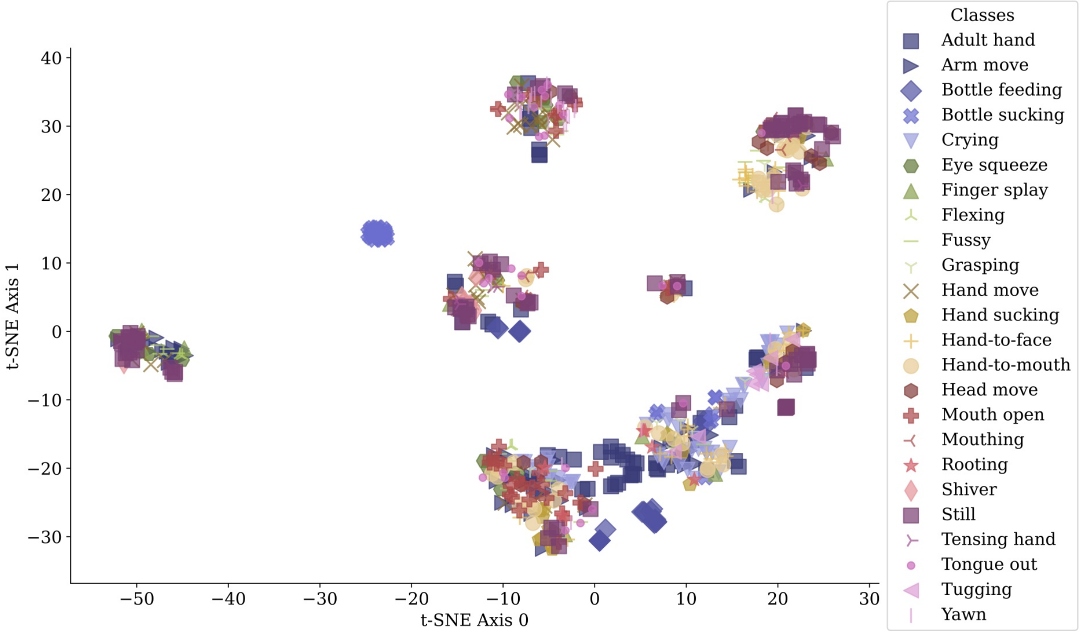


Figure 5.6: V2 — t-SNE visualisation of the SLAPI dataset using the MoViNet A2 model backbone, after applying optical flow preprocessing. It visualizes the high-dimensional data in a two-dimensional space, revealing patterns between data points. Each marker represents a video snippet, while classes are separated by marker style and color. The proximity of the points reflects the similarity of their feature representation when reduced to a 2D space.

flow per pixel, and the third channel for the gray-scale frame. This ablation condition presented a stark contrast to the baseline experiment, and performance declined for all metrics. Validation and test accuracy decreased to .13 and .14 respectively. Furthermore, precision, recall, and F_1 score were similarly reduced to .07, .16, and .08 respectively. The t-SNE visualization in Figure 5.6 did not yield encouraging results either, showing a similar effect of the source video on the representation. These results suggest that the more complicated preprocessing step introduced some form of noise in the representation, or the model could not adapt from the RGB pretraining to optical flow representations. The number of training epochs was not limited, but validation loss did not improve after the third epoch. Metrics were computed using the weights from this epoch.

The final data modification condition V3 employed leave-one-out cross validation using the baseline approach. This condition was created to gain better insight in the significance of each infant to the classification of their cues. Table 5.6 presents the results of this leave-one-out cross validation tasks, where one infant is left out of the training data and used as test set. To ensure fair comparison between the infants, classes that were not exhibited by all infants are removed from the dataset. Consequently, the dataset is reduced to four classes due to the limited behaviors exhibited by infant 74. After excluding this infant from the dataset, 9 classes remain. For each class, the precision, recall, and F_1 score are reported. The last row presents the mean performance across all classes for an infant.

From the table, it can be gathered that performance varies significantly across the infants. Generally, mean performance is worse for infant 50 across all metrics (P=.23, R=.19 and F_1=.18). This suggests that the way this infant displays its cues varies from the way the other infants do. For instance, when looking at the cue "*Still*", the precision score is much lower than for the other infants. In similar vein, the classes "*Finger splay*", "*Fussy*" and "*Yawn*" are only correctly predicted for infant 79, which suggests that the yawning does not generalize well to the other three infants. The reverse is true for "*Hand-to-*

Class	**Precision**				**Recall**				F_1			
Infant ID	*50*	*79*	*95*	*70*	*50*	*79*	*95*	*70*	*50*	*79*	*95*	*70*
Adult hand	.94	.76	.67	.15	.56	.87	.95	1.0	.70	.81	.78	.27
Finger splay	.00	.12	.00	.00	.00	1.0	.00	.00	.00	.22	.00	.00
Fussy	.00	.33	.00	.00	.00	.09	.00	.00	.00	.14	.00	.00
Hand-to-face	.07	.00	.54	.38	.25	.00	.86	.70	.11	.00	.67	.49
Hand-to-mouth	.00	.50	.00	1.0	.00	.08	.00	.19	.00	.14	.00	.32
Head move	.02	.00	.00	.33	.25	.00	.00	.17	.04	.00	.00	.22
Still	.67	.94	.85	.86	.55	.76	.53	.78	.60	.84	.64	.82
Tongue out	.33	.00	.18	.00	.11	.00	1.0	.00	.16	.00	.31	.00
Yawn	.00	.20	.00	.00	.00	.50	.00	.00	.00	.29	.00	.00
Mean	.23	.32	.25	.30	.19	.37	.37	.31	.18	.27	.27	.23

Table 5.6: V3 — Leave-one-out cross-validation evaluation metrics displayed in terms of Precision (P), Recall (R), and F_1 score for each class across four different infants (ID: 50, 79, 95, 70). The respective infant's data was held out as a test set for evaluation. This table provides an insight into the model's ability to generalize across different infants, highlighting variability in performance metrics across different classes and infants. This approach allows for assessing the extent to which the individual infant's data influences the model's performance.

face", which is only correctly predicted in infant 79. When infant 95 is the test set, the precision score (.25) is significantly lower than the recall (.37). This could mean that the model tends to over-predict the classes "*Adult hand*", "*Hand-to-face*", "*Still*" and "*Tongue out*". This over-prediction could result in an increased number of correct detections but also lead to a higher rate of misclassification for these classes. As for infant 70, it is noticeable that for the classes "*Hand-to-mouth*" and "*Still*" the precision (1.0 and .86 respectively) is relatively high, but the recall is lower (.19 and .78 respectively). The model is conservative in predicting these classes, but when it does, it is usually correct. The reverse is true for "*Adult hand*" (P=.15 and R=1.0) and "*Hand-to-face*" (P=.38 and R=.70), where it identifies most of the occurrences in the test set, but at the cost of more misclassifications. This variability in performance patterns across infants may indicate the different behavior profiles of infants. This highlights the influence of individual differences among infants and the specific cues on the model's predictive capability and capacity for generalization. This underlines the inherent complexities in accounting for individual differences. To visually inspect performance per infant, the confusion matrices for each infant are shown in Figure C.1, reported in Appendix C.

5.4.2 Results Component Modifications

The results of the ablation study which evaluated the effect of different model components on performance are depicted in Table 5.7. Despite proven effectiveness in the literature and extensive use in implementations, certain components like regularization, causal convolutions, and backbone fine-tuning for feature representations were not included in the final model for the baseline experiment by the genetic search for hyperparameters. It was hypothesized that the model components may not have been selected due to memory and time constraints on the genetic search process. The components are implemented in these ablation studies and tested on both the baseline and FSL approaches to see if performance can improve.

Exp.	**Version**	**Val. Acc.**	**Acc.**	**P**	**R**	F_1
Baseline	V4	.04	.02	.02	.05	.01
	V5	.17	.12	.04	.11	.04
	V6	.14	.14	.10	.13	.06
	V7	.04	.05	.01	.06	.02
	V8	.25	.21	.18	.21	.14
	V9	.36	.31	.23	.27	.20
	V10	.22	.23	.13	.24	.14
FSL	V4	.02	.04	.00	.04	.00
	V5	.14	.15	.04	.08	.04
	V6	.05	.04	.02	.05	.03
	V7	.11	.10	.11	.11	.05
	V8	.07	.07	.00	.04	.01
	V9	.16	.18	.14	.21	.14
	V10	.03	.00	.00	.01	.00

Table 5.7: V4-V10 — Comparison of evaluation metrics displayed in terms of Precision (P), Recall (R), and F_1 score for different ablation conditions (V4 - V10) of the baseline and FSL approach.

In the baseline experiment, version V9 demonstrates the highest performance across all metrics: validation accuracy (.36), accuracy (.31), precision (.23), recall (.27), and $F1$ (.20). This version uses a frozen backbone training strategy, enabled causal convolutions and no regularization, suggesting that these configurations were the most effective for this specific task within the baseline experiment. Though superior in this context, this highest performing version falls short of the baseline experiments (*validation acc.*=.35(.03), *acc.*=.30(.02), P=.28(.02), R=.30(.02) and F_1=.25(.02)). Conditions V8 and V10 also perform relatively well in terms of accuracy (.21 and .23 respectively). However, precision, recall and F_1 did not score well. In both experiments the model backbone is trainable and regularization is disabled, but in V8 the causal convolutions are enabled as well.

As for synergistic effects, no noteworthy patterns arise from the table. The observed effects of the three variations appear to be independent of other component settings. For instance, the absence of regularization (as seen in V8, V9, and V10), seems to generally result in better model performance. This indicates that the L2 regularization is too restrictive for this task and might be hindering the model's ability to capture the complexity of the data. A trainable model backbone (V4, V6, V8, V10) also results in worse performance compared to its counterparts. Most likely, the model is unable to transfer the properties of the limited data to the extremely large parameter space. When using causal convolutions, however, it is notable that model performance severely deteriorates when regularization is applied (V4, V5). Without regularization (V8 and V9), the models are competitive with the baseline approach. This could suggest that the model is benefiting from the consideration of only unidirectional temporal dependencies in the data, when it is not restricted by regularization.

In the FSL, the best performing ablation version is again V9 across all metrics: validation accuracy (.16), accuracy (.18), precision (.14), recall (.21), and $F1$ (.14). However, these results are significantly worse compared to the main FSL experiment (*validation acc.*=.26(.03), *acc.*=.26(.03), P=.23(.03), R=.26(.03) and F_1=.20(.03)). Especially the use of a trainable backbone in V4, V6, V8 and V10 debilitated the model's performance. Ablation version V8 shows most promise in an otherwise disappointing set of models, with validation accuracy, accuracy and recall being higher than .01 (.07, .07 and .04 respectively). No noteworthy synergistic effects arise from these experiments.

The main conclusion that can be drawn from this ablation study is that the model found by the genetic search was indeed the best model for this dataset. Model performance deteriorates slightly when enabling causal convolutions and significantly when using L2 regularization in the baseline approach. In the FSL approach training the model backbone appears to prevent any meaningful learning from taking place.

5.4.3 Results Active Learning Modifications

Under ablation condition V12, it is tested whether stratified maximum probability sampling could enhance the localization of underrepresented classes, by comparing the approach to a more standard uncertainty sampling approach. Other than that, the same settings were used as in the AL experiment. Using uncertainty sampling, validation accuracy (.36 versus .38), accuracy (.34 versus .35), precision (.26 versus .25), recall (.30 versus .30) and F_1 (.26 versus .26) were similar to stratified sampling. This suggests that uncertainty sampling does not affect model performance on the whole or in underrepresented classes, when compared to stratified sampling.

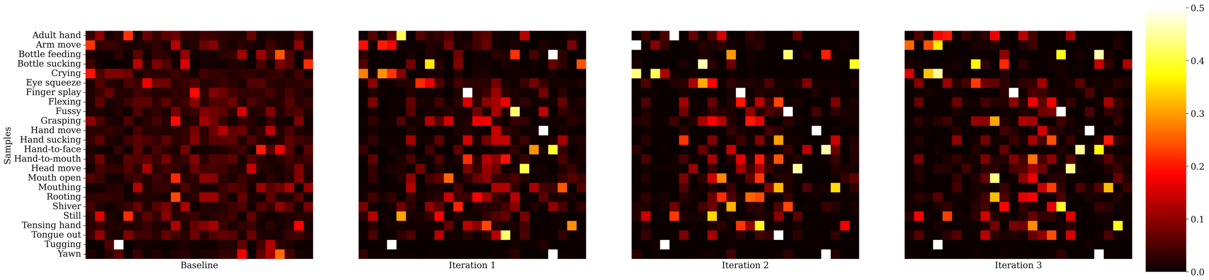


Figure 5.7: V12 — A series of heatmaps shows the evolving class probabilities for 24 samples across AL iterations using uncertainty sampling. The y-axis orders samples by class, while the x-axis represents class probabilities. The heatmaps chronologically illustrate the progression of these probabilities.

Table C.1 in Appendix C shows the labeling of snippets during the iterations. The class “Still” emerges as the class that is most often selected identified by the oracle, with a total of 75 samples. Other relatively common classes were “*Arm move*” and “*Head move*”, with a total of 23 and 15 samples. In the third AL iteration, more varying classes are identified by the oracle, being “*Crying*”, “*Eye squeeze*”, “*Finger splay*”, and “*Yawn*”.

Figure 5.7 presents heatmaps that represent the evolution of class probabilities for 24 samples through the AL iterations using uncertainty sampling. After the first iteration, in which 50 samples are labeled and the model is trained for another 3 epochs, the distribution of the probabilities becomes more sparse. This indicates that the model’s predictions are becoming more certain, with the classification probability centering around fewer classes. After the second iteration, the learned parameters are consolidated as the remaining uncertainty is further reduced. Between the second and final iteration, there are no notable changes in the probabilities. This implies that the model’s learning has plateaued at this stage, without additional meaningful insights or understanding gained from the final iteration, despite the inclusion of new instances. As in stratified sampling, it appears the probabilities convergence mostly around classes that are incorrect.

6 Discussion

This study investigated the use of computer vision for classifying infant hunger and feeding discomfort cues. The aim was to determine the capabilities of different machine learning approaches in classifying preterm infant behavioral cues on the NICU. The results from this study build towards a system that could facilitate cue-based feeding approaches due to reduced monitoring burden, to the benefit of the health of preterm infants admitted to the NICU. For this investigation, 3D CNN MoViNets designed by Kondratyuk et al. [2021], were tested in three distinct approaches to determine their capability in infant monitoring. Other than the standard fully supervised approach, FSL and AL were chosen due to their ability to operate effectively with limited data. This is a challenge inherent to machine learning approaches, where larger datasets are typically preferred, but not available in this setting.

6.1 Findings and Interpretation

The main research question, "*Can we detect and classify infant hunger and feeding discomfort cues in preterm infants using machine learning?*", was constructed in order to evaluate the current capabilities and limitations of state-of-the-art methods for infant monitoring. In order to comprehensively answer the main research question, several sub-questions were postulated that addressed the impacts of FSL and AL approaches. These strategies propose different solutions to the limited availability of training data, thereby highlighting the constraints that currently exist in a standard machine learning approach and how these may be overcome. The third sub-question was constructed to evaluate the impact of individual differences on classification.

The first two sub-questions were addressed through the main experiments, training the MoViNet model via the respective pipelines. The third sub-question was addressed using leave-one-out cross validation on the baseline approach. In this section, the findings from this study are reviewed, interpreted, and placed in the wider context of computer vision, motion analysis and infant monitoring.

6.1.1 Sub-questions

The first sub-question, "*SQ1: Can we detect and classify infant hunger and feeding discomfort cues in preterm infants with a model trained using few-shot learning?*" originated from the hypothesis that the meta-learning phase could optimize search path through the parameter space, thereby guiding it towards quicker, more effective convergence during the final testing phase. This approach had proven successful in similar applications such as infant facial recognition systems distinguishing infants from adults [Atallah et al., 2022], and human motion prediction [Gui et al., 2018]. The application of meta-learning to intricate facial details and other human actions using computer vision suggested that FSL could result in improved performance over a standard machine learning approach. However, contrary to these expectations and a volume of literature supporting the FSL framework when working with small datasets [Finn et al., 2017; Romanov et al., 2021; Jamal and Qi, 2019; Mohammadi et al., 2019], the evidence showed decreased performance across all metrics.

The unexpected finding may be explained by challenges in translating parameters learned from the source to the target domain. Finn et al. [2018] found that a critical challenge for meta-learning is task-ambiguity. This holds that even when weights can be successfully tuned in the meta-learning phase, that the target domain is too ambiguous or unrelated to make use of these weights. This problem is amplified in small datasets, a common feature in contexts where meta-learning is employed. In an attempt to combat this issue, the meta-tasks were sampled from the UCF101 dataset (Section 3.1.8), ensuring the same number of outcome classes in each randomly sampled meta-task to mirror the target domain's ambiguity. This dataset is particularly suited for transfer learning or meta-tasks due to its high degree of intra-class variance. This variance acts as a rich learning source for the model, allowing it to learn how to generalize its parameters to another domain due to the robustness required for the variations within the classes. However, it appears that the SLAPI dataset's inherent ambiguity and noise, a point that is frequently discussed in this thesis, remained a significant challenge. Consequently, the FSL meta-learning failed to acquire a model that improved performance in this setting.

The survey of Hospedales et al. [2022] further raised the concern that a single model initialization may not be sufficient to generate models that fit a wide range of tasks. They suggested that multiple initializations and mixtures of these models could yield better performance. Their survey indicates that task complexity is an important factor in determining the success of FSL. It highlights that previous studies generally achieved success with simpler tasks, particularly those in which the source and target domains exhibited a high degree of task relatedness. The classification problem in this thesis was relatively complex. Although all classes concerned infant behaviour, the variance in behaviours as well as variance introduced by the source videos could have contributed to the disappointing results. Furthermore, performance in this set up may have suffered due to the discrepancy between meta-learning and meta-testing tasks, making the model unable to effectively prepare its parameters for the meta-testing phase. This is often inevitable due to the nature of small datasets and setups. However, it often proves detrimental to performance. For instance, Guo et al. [2020], found that ImageNet tasks do not generalize well to specialist domains such as medical images, which has proven a struggle in this thesis too. They further demonstrated that FSL approaches were outperformed by simple fine-tuning approaches such as standard transfer learning, noting that accuracy is correlated with similarity between source and target domains. These studies collectively confirm the finding that a meta-learning FSL approach was not successful in performing cue classification for preterm infants, despite the approach being successful under more favourable conditions.

The second sub-question, "*SQ2: Can we detect and classify infant hunger and feeding discomfort cues in preterm infants with a model trained using active learning?*" was constructed based on the presumption that underrepresented classes and informative samples should be sought out for improved performance. It was previously shown to be effective by Yang et al. [2015] in a multi-class setting, when applied to the large standardized YouTube dataset. Though AL has not been previously applied to infant monitoring or similar tasks, it seemed especially suited for this study's conditions, where only limited and unbalanced data was available. An additional benefit of the AL approach was the potential for a paradigm shift identified by Budd et al. [2021]. When successful, AL could feature easier labeling which would improve model performance over time. It was found that this hypothesis was somewhat supported by the evidence, which showed comparable performance to a standard machine learning approach.

This outcome aligns with literature showing that AL was successful in multi-class classification tasks [Joshi et al., 2009; Lorbach et al., 2019] and a setting of action localization when labeled data was scarce [Heilbron et al., 2018]. However, the results did not entirely support the hypothesis. No improved performance was observed compared to a standard machine learning approach, and it appears as though stratified sampling did not improve the detection of underrepresented classes. The work by Stumpf et al. [2014] had previously shown that most sampling strategies do not consider spacial constraints in the representation, leading to a distribution of training instances that may be unfavourably compact or sparse. By applying a standard uncertainty-based sampling strategy it was expected that majority classes would be over-sampled due to their prevalence in the dataset and more varying representations. When using stratified sampling it was expected this would yield beneficial results for underrepresented classes. This was corroborated by Rougier et al. [2016] who showed that using stratified sampling over regular sampling leads to lower labeling costs by requiring fewer samples to achieve similar performance. Contrastingly, this study revealed no discernible advantage of stratified sampling over standard uncertainty sampling. This can be explained by the possible absence of the underrepresented classes in the unlabeled pool. Additionally, the stratified sampling strategy employed in this thesis attempted to find underrepresented classes by finding the samples with the highest probability of belonging to such a class, not necessarily the highest probability overall. Combined, these factors explain the lacking performance of AL in underrepresented classes. Another discouraging result from the experiment was the lack of improved performance despite the 150 additional samples. This outcome suggests that the samples added did not significantly enhance the learning process, contrary to what was hypothesized. Since it was established that the source video greatly affects the feature representation, it is possible that the data from the unlabeled pool did not improve performance as this source video was not included in the test set. When the samples from the AL were used as the training set, and the original training set was used as the test set, there was a significant reduction in performance. This implies that the AL samples do not generalize to the dataset effectively. Introducing new instances in the training set may only hinder performance as it obfuscates the relation between the train and test set. This could also explain why the model does not seem to correct the incorrect predictions in the 24 samples that were set aside. Despite this, AL still successfully learned parameters and exhibit convergence, as demonstrated by the probability analysis.

Despite these sub-optimal results, a benefit of the pipeline is the labeling framework, appraised by Lowell et al. [2018] and Budd et al. [2021]. They discussed the implementation and argued that the human-in-the-loop could insert expert knowledge into models that require it for medical image analysis, for instance. They assert that, by incorporating confidence scores into the oracle, a more *human interpretable* metric is available through the application of AL. This approach then aligns more with the conceptual frameworks of doctors and other medical professionals, thereby making this approach more suitable for those who might use it. So, AL served as a useful labeling tool and allowed for model fine-tuning. Therefore, even considering the setbacks, the evidence still supports the hypothesis that AL is suitable for cue classification and data acquisition.

The third sub-question was formulated as "*SQ3: Do individual differences in infants and cues affect automated detection and classification of behavioral cues?*". Based on prior studies revealing variations between infants in terms of the onset, occurrence, intensity, and form of their behaviors [Thoman and Whitney, 1990; Frischen et al., 2007; Claessens et al., 2011], it was hypothesized that these individual differences could significantly influence the classification results. For instance, research showed that bottle-feeding outcomes are greatly affected by the individual differences in satiation cues by infants, highlighting differences in urgency, impulsivity and negative temperament [Ventura and Mennella, 2017]. The findings of this study support the hypothesis, confirming that individual differences in behavioral cues exist and impact classification.

Throughout the study, significant fluctuations in model performance were observed across different classes and among different infants. These variations persisted when compared with the baseline model, highlighting the importance of individual differences among infants and the specific exhibited cues. However, this study found that under certain conditions the impact of individual differences on classification could be mitigated. When pooling the data of the infants, the effects of individual differences were somewhat obscured, as evidenced by improved performance in the baseline approach compared to the leave-one-out cross validation ablation. Creating a sufficiently large pool of infants the dataset could diminish the impact of individual differences on performance. Further, when new infants are admitted to the NICU, adding training data of that infant to the pool may result in improved performance as it allows the model to learn the infant's unique characteristics. The snippet-based approach removes the effect of onset and frequency in the classification process. As a result, these characteristics of the cues are of no impact on the model performance but do still exist in the infant. Therefore, the implications of cue onset and frequency must still be considered when using the system. For example, in the untrimmed videos that were profiled, the frequency and onset of different cues may be determined by inspecting the chart. Currently, such an approach is preferred as the limited literature on infants cues means that it remains difficult to draw definitive conclusions on the basis of onset and frequency using a machine learning approach.

The influence of visual differences arising from the source videos may obscure the direct effects of behavioral differences, preventing strong conclusions. Nevertheless, the leave-one-out cross validation ablation demonstrated that differences vary with cues, and were not identical across infants (or source videos). This implies that the variations in infant behavior, at least to a some degree, have contributed to these differences, impacting the model's performance. In this thesis, the effects of onset and frequency were removed by the snippet-based approach, and should be studied further to document its meaning. The current observed discrepancies emphasize the complexity of accounting for individual differences, also underlining the need for further research to improve model performance.

6.1.2 Main Question

The main hypothesis concerned the broad feasibility of machine learning approaches for monitoring infant behaviour. It was hypothesized that, under the constraints in data availability and quality, promising outcomes were plausible. This was supported by the work of Sun et al. [2019] and Sun et al. [2021], who explored classification of infant states and specifically discomfort. After addressing the sub-questions to this main research question, it can be concluded that the sub-questions support this hypothesis. Not only were the obstacles of individual differences, limited data, and quality of labels overcome, the approach yielded improved performance over majority class voting and weighted guessing, showing that meaningful features were learned. After having reviewed the performance of the baseline approach on a model- and cue-level it was evident that the results of this thesis allow for a monitoring approach that can reduce the burden on current NICU monitoring. Individual and visual differences in cues are an important property of the data but do not prohibit machine learning approaches to this problem. The use of AL further enhances the monitoring, including a data generation framework and possibly enhancing underrepresented classes.

In a comparative analysis, with the standard supervised approach as baseline, it turned out that FSL method revealed a noticeable drop in performance. Evidence from the genetic search suggests the FSL approach was not suited to the classification problem, as can be concluded from the small number of shots and high learning rate. These factors limit the impact of the meta-learning phase by limiting the samples available for learning and enabling rapid adjustment of parameters via high learning rates. It is apparent that the AL is more suited to the medical domain, solving the issue of small datasets more effectively. It also aligns more closely with the conceptual framework medical professionals operate in and it simply performs better. This discrepancy can largely be attributed to the inability of FSL to generalize the learned features from the source domain. Conversely, the AL approach maintained performance despite not finding informative samples for learning, which indicates its robustness even when results are not as beneficial as expected.

All in all, the main research question is answered positively. This means that this study confirms the assertion that infant behaviour can be studied using machine learning approaches, that were not previously applied in this specific setting for hunger and feeding discomfort cue detection. This was achieved using a limited dataset, which yielded promising results. Although classification metrics are not sufficiently high for definitive conclusions, the models developed in this thesis can help in delineating such behaviours. These systems are designed with the intention of aiding nurses by reducing the monitoring burden. To this end, ground truths are not an essential requirement. The burden of monitoring can be alleviated by interpreting the variations in behaviors, regardless of the accuracy of their classification. Due to the high precision scores in outcome classes such as "*Still*" and "*Bottle feeding*" and "*Adult hand*" conclusions may be drawn from their predictions and its patterns. These conclusion can then serve to support nurses. With facilitating cue-based feeding as the aim, extensive monitoring is required to make sure infants do not grow excessively hungry, and are not left in pain. This study presents an initial step towards developing a system that can determine feeding moments. In its early stages, the system can alert nurses to attend based on the predicted actions, even if the predictions are not perfect.

6.2 Limitations

The main limitations pertain to the SLAPI dataset. This is not unexpected as the quality of the data is highly important to model performance. Although the limited quantity of labeled data across classes was known from the start of this research, and modifications were made to combat these issues, it was still a considerable drawback that impacted performance

First off, there were issues with the consistency of the recorded videos. Despite attempts to maintain uniform camera positions using the setup, not all videos adhered to this configuration. Particularly in videos recorded in the early stages of the SLAPI study, which displayed more discernible cues. Cameras were manually positioned on the NICU bed, resulting in some variability. It has long been common knowledge that viewpoint consistency is important in model-based computer vision system, as was illustrated by Lowe

[1987]. As they formulated it, applying a fixed point of view simplifies the classification problem, as the location of the cues of interest is no longer a variable to consider, and features may be dedicated to the action recognition.

Another source of noise is the lack of consistency in surroundings further compromising the dataset. Factors such as the bed, infant clothing, and medical equipment varied significantly across videos. Ideally, all subjects would be recorded under similar circumstances, or in large enough numbers that this becomes natural variation. However, this was unachievable given the current phase of the recording setup. These combined imperfections likely contributed to different source videos being recognized from the feature representation using t-SNE plot in Figure 3.1. Without these noisy circumstances, it is more likely that cues are clustered by their correct class, rather than the untrimmed source video they originate from.

Another limitation concerned the annotations. Although effort was put into the consistency of the annotations, they were performed by research interns who previously had no experience with interpreting infant behavior. Subjective labels are problematic when they introduce bias into the dataset, as outlined Miceli et al. [2020]. They argued the annotation process starts as soon as the annotator forms needs and expectations surrounding the data. As was the case in this study, there were no other annotators available for this dataset, so common annotation validation tools such as inter-rater reliability were not available. This potentially allowed the introduction of bias.

Finally, the snippet-based approach presented its own challenge. This approach disrupted the original sequence of the data, thereby omitting information about transitions, such as the onset and stopping of actions. While this approach did capture some variation, with snippets potentially capturing the beginning, middle, or end of a behavior, it did complicate the process of drawing conclusions about the increasing intensity of hunger or feeding discomfort. It was shown that cues are more frequent and more intense as the infant gets more hungry [Whetten, 2016] or is in more discomfort [Morison et al., 2003]. They also showed that infants display a larger variation as the pain becomes more intense. However, profiling an untrimmed video through predictions of consecutive snippets still reveals pattern in frequency and variation, although not inherent in the knowledge of the model. Only the transitions in cues are more difficult to capture and may be obscured by more obvious cues.

6.3 Validity and Generalizability

The validity of this study may be impacted by the absence of medical validation in certain aspects of the research, such as hunger and feeding discomfort cues. Whilst these terms have clear common-sense definitions, they are not adequately validated in the literature [Ludwig and Waitzman, 2007]. Therefore, it is challenging to interpret cues in terms of these infant states. While there are validated scales to identify infant states such as pain or readiness for oral feeding [Fujinaga et al., 2013], these scales fail to adequately capture the nuances of infant hunger states and potential subsequent discomfort during feeding [Paul et al., 2014]. Due to this complication, cues which are ambiguous, or cues which span multiple states simultaneously, are challenging to interpret. To mitigate this issue in this study, particular definitions were not ascribed to the observation of behavioral cues wherever possible. Instead, only factual descriptions of the observed behaviors were used, avoiding subjective interpretations.

The generalizability of this research mainly concerns the leave-one-out cross validation ablation study and the extent to which the findings are applicable to different infants. From the ablation study, it was apparent that individual differences between infants certainly do affect classification. Nevertheless, similar patterns occurred in majority classes across all infants despite these individual differences, leading to the conclusion that there is definitely sufficient consistency between infants to apply the system to infants it was not trained on. Another potential issue is that diversity of the infants may not have been sufficient to translate to different settings. The subject pool was not sufficiently varied, which raises concerns about the study's susceptibility to a commonly occurring bias known as *whiteness* in medical research and education [Zaidi et al., 2023]. This bias, rooted in the overrepresentation of certain demographic groups, can compromise the generalizability of the research outcomes to broader, more diverse infant populations. Finally, a more accurate assessment of model generalizability could be achieved by ensuring that the same videos are not incorporated in both the training and testing datasets. As suggested by the t-SNE visualizations, the model may struggle to generalize content from videos that deviate significantly from the ones it was trained on, as the video's currently impact the feature representation of cues that are trimmed form it.

Overall, the validity of the results in this study relies on the cautious interpretation of infant behaviour, contributing to the robustness of the findings. Generalizability, on the other hand, is more suspect as it may be compromised twofold: by the potential *whiteness* and compromised feature representations, leading to poor performance in the ablation study. Thus, it is essential to consider these elements when evaluating the findings and their broader applicability.

6.4 Future Directions

From the findings and limitations, interesting paths for future research were identified. Firstly, this thesis attempted to address the limited data in the model domain by introducing AL and FSL. However, it would be interesting to see how this issue could be addressed in the data domain. Yun et al. [2020] reported that typically image augmentation techniques are appplied to videos on a frame-to-frame basis, like flipping, cropping and background subtraction. To apply augmentation to temporal information, sub-sampling video frames is commonly used as augmented representation. However, there is a lack of studies investigating how more robust features can be learned from video data for action localization with acceptable generalizability.

A further research direction concerns domain adaptation. Due to the novelty this research, only few papers have been published in which a machine learning approach has been applied to the NICU. One such study is the research by Sun et al. [2021] who performed camera-based discomfort detection on infants admitted to the NICU. As a result, there are minimal suitable source domains available for the target domain of hospitalized infants. Due to the importance of suitable source tasks for transfer learning, this suggests that domain adaptation causes a performance bottle neck in this study. Wang and Deng [2018] found that deep domain adaptation can be used to address the lack large volumes of labeled training data. However, they single out task relatedness as an important consideration for domain adaptation. Wilson and Cook [2020] state that tasks are sufficiently related if they use the similar features during classification, Xue et al. [2007] found that domains should have feature vectors in near each other in the feature space, while Ben-David and Schuller [2003] suggest that tasks are related if they can be drawn from a fixed probability distribution (given appropriate transformations). Due to the significance of task relatedness, future research could focus on finding more suitable pretraining tasks for target domain of preterm infants.

Another promising strategy for model improvement is the use of a multi-modal approach, which has been proven successful in different domains [Bayoudh et al., 2022; Maragos et al., 2008; Baltrušaitis et al., 2017]. Other than simply using video input, modalities such as heart rate, oxygen saturation and audio could be integrated into the information used for classification. These three modalities currently serve as an important measure of infant health. For example, specific heart rate characteristics are frequently used indicators of infant health [Hicks and Fairchild, 2013]. Clear thresholds have been established for oxygen saturation levels, identifying when infants are at risk [Stenson et al., 2013]. It has been shown how crying melodies differ for pain, hunger or normal cries [Rothgänger, 2003]. Therefore, the inclusion of these modalities warrants further investigation.

Furthermore, it is recommended that future studies look at the generalizability of such results to wider infant populations by investigating individual differences further. This could also involve the inclusion of infants of varying ages, different stages of development or health conditions. Such studies would not only validate the conclusions drawn from this research, but also provide a wider understanding of how classification is affected by individual differences. Opportunities for enlarging the data pool could involve a more systematic approach in data collection, such as coordinating with larger research groups to collect data. Existing data could also be exploited to generate more training data. The approach by Yang et al. [2022] serves as an example as example in combining synthetic data with real data to achieve improved pose recognition.

The explainability of models is becoming increasingly important. As decision-making processes need to be rationalized, the *black box* approach cannot be used to explain decisions. This necessitates further investigation into how the decisions made by increasingly complex models can be effectively explained to those working with automated monitoring systems. In this respect, AL is a suitable option. It fits within the conceptual framework familiar to medical professionals, presenting instances that meet a specific selection criterion. This way, professionals gain a better understanding of the knowledge gaps in the systems and how to interpret this information. This aligns with the findings of Ren et al. [2021], who highlighted the importance of expert knowledge in the system applied in a medical to prevent potential degradation of quality, a concern raised by Lowell et al. [2018] and Budd et al. [2021].

Finally, although beyond the scope of this thesis, future research could validate hunger and feeding discomfort cues. One way this could be achieved is by tracking feeding times and noting how the feeding is performed, and if the food is accepted fully. Ethical considerations prohibit experimenting with more excessive hunger states by delaying feeding, so these studies must be meticulously planned. This validation must also consider that behavioral cues vary depending on the infant's state. For example, an open mouth may have different meanings when the infant is in quiet sleep versus when awake. Future studies could potentially build upon the findings of this research by incorporating these differing states in a more comprehensive manner. This could lead to more accurately attributing behavioral cues to hunger or other states. As was laid out previously, a more profound understanding of infant cues can contribute to increased usability of monitoring systems, elevating the overall quality of care.

7 Conclusion

This thesis presents a comprehensive examination of various machine learning approaches, specifically applied to a novel context that holds significant potential benefits for the health of preterm infants admitted to hospitals so early in their lives. The motivation for this work was the development of a monitoring system that enables cue-based feeding in NICUs, an advancement that can profoundly impact preterm infant health outcomes.

While the outcomes presented in this thesis may not yet possess the robustness required for immediate application in such a monitoring system, the findings are nonetheless promising. A noteworthy result is the demonstrated ability of standard machine learning techniques to construct meaningful profiles from untrimmed data through snippet classification. This ability can be readily leveraged for several applications such as care quality inspection and as an auxiliary monitor for infant restlessness, supplementing traditional heart rate monitors.

The envisioned monitoring system does not require flawless performance due to the cyclical nature of hunger cues and feeding discomfort. These cues are repeated over time and vary in intensity and frequency, allowing for recognition patterns even without perfect system performance. This aspect might also bridge individual differences among preterm infants since no single cue is definitive.

Overcoming challenges encountered during this study, particularly those related to data quality, would mark a significant step forward. When medical validation of preterm infant behavioral cues is more comprehensively established, the envisioned system could become a reality. Consequently, the goal of introducing cue-based feeding in the NICU will move one step closer to fruition.

References

Alibrahim, H. and Ludwig, S. A. (2021). Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. In *2021 IEEE Congress on Evolutionary Computation (CEC)*, pages 1551–1559, Kraków, Poland. IEEE.

Atallah, R. R., Kamsin, A., Ismail, M. A., and Al-Shamayleh, A. S. (2022). Neural network with agnostic meta-learning model for face-aging recognition NN-MAML for face-aging recognition. *Malaysian Journal of Computer Science*, 35(1):56–69. Number: 1.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2017). Multimodal Machine Learning: A Survey and Taxonomy. arXiv:1705.09406 [cs].

Bayoudh, K., Knani, R., Hamdaoui, F., and Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970.

Ben-David, S. and Schuller, R. (2003). Exploiting Task Relatedness for Multiple Task Learning. In Schölkopf, B. and Warmuth, M. K., editors, *Learning Theory and Kernel Machines*, Lecture Notes in Computer Science, pages 567–580, Berlin, Heidelberg. Springer.

Bertelle, V., Sevestre, A., Laou-Hap, K., Nagahapitiye, M. C., and Sizun, J. (2007). Sleep in the Neonatal Intensive Care Unit. *The Journal of Perinatal & Neonatal Nursing*, 21(2):140.

Bertinetto, L., Henriques, J. F., Valmadre, J., Torr, P., and Vedaldi, A. (2016). Learning feed-forward one-shot learners. *Advances in neural information processing systems*, 29.

Blauer, T. and Gerstmann, D. (1998). A Simultaneous Comparison of Three Neonatal Pain Scales During Common NICU Procedures. *The Clinical Journal of Pain*, 14(1):39.

Budd, S., Robinson, E. C., and Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062.

Bukschat, Y. and Vetter, M. (2020). EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach. *arXiv preprint arXiv:2011.04307*.

Cao, G., Liu, J., and Liu, M. (2022). Global, Regional, and National Incidence and Mortality of Neonatal Preterm Birth, 1990-2019. *JAMA Pediatrics*, 176(8):787.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299.

Chambers, C., Seethapathi, N., Saluja, R., Loeb, H., Pierce, S. R., Bogen, D. K., Prosser, L., Johnson, M. J., and Kording, K. P. (2020). Computer Vision to Automatically Assess Infant Neuromotor Risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(11):2431–2442. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., and Zhang, L. (2020). HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. arXiv:1908.10357 [cs, eess].

Cherian, A. and Gould, S. (2019). Second-order temporal pooling for action recognition. *International Journal of Computer Vision*, 127:340–362. Publisher: Springer.

Chrupcala, K. A., Edwards, T. M., and Spatz, D. L. (2015). A Continuous Quality Improvement Project to Implement Infant-Driven Feeding as a Standard of Practice in the Newborn/Infant Intensive Care Unit. *Journal of Obstetric, Gynecologic & Neonatal Nursing*, 44(5):654–664.

Claessens, S. E., Daskalakis, N. P., van der Veen, R., Oitzl, M. S., de Kloet, E. R., and Champagne, D. L. (2011). Development of individual differences in stress responsiveness: an overview of factors mediating the outcome of early life experiences. *Psychopharmacology*, 214:141–154. Publisher: Springer.

Clark, A., Donahue, J., and Simonyan, K. (2019). Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*.

Du, X., Li, Y., Cui, Y., Qian, R., Li, J., and Bello, I. (2021). Revisiting 3D ResNets for video recognition. *arXiv preprint arXiv:2109.01696*.

Embleton, N. D. (2013). Early Nutrition and Later Outcomes in Preterm Infants. *Nutrition and Growth*, 106:26–32. Publisher: Karger Publishers.

Ertugrul, I. O., Jeni, L. A., Ding, W., and Cohn, J. F. (2019). Afar: A deep learning based tool for automated facial affect recognition. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–1. IEEE.

Fanaro, S. (2013). Feeding intolerance in the preterm infant. *Early Human Development*, 89:S13–S20.

Farnebäck, G. (2003). Two-Frame Motion Estimation Based on Polynomial Expansion. In Bigun, J. and Gustavsson, T., editors, *Image Analysis*, Lecture Notes in Computer Science, pages 363–370, Berlin, Heidelberg. Springer.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. arXiv:1703.03400 [cs].

Finn, C., Xu, K., and Levine, S. (2018). Probabilistic Model-Agnostic Meta-Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Frischen, A., Bayliss, A. P., and Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4):694. Publisher: American Psychological Association.

Fry, T. J., Marfurt, S., and Wengier, S. (2018). Systematic Review of Quality Improvement Initiatives Related to Cue-Based Feeding in Preterm Infants. *Nursing for Women's Health*, 22(5):401–410.

Fujinaga, C. I., Moraes, S. A. d., Zamberlan-Amorim, N. E., Castral, T. C., Silva, A. d. A., and Scochi, C. G. S. (2013). Clinical validation of the preterm oral feeding readiness assessment scale. *Revista latino-americana de enfermagem*, 21:140–145. Publisher: SciELO Brasil.

Fuller, B. (1996). Meanings of discomfort and fussy-irritable in infant pain assessment. *Journal of Pediatric Health Care*, 10(6):255–263.

Gong, X., Li, X., Ma, L., Tong, W., Shi, F., Hu, M., Zhang, X.-P., Yu, G., and Yang, C. (2022). Preterm infant general movements assessment via representation learning. *Displays*, 75:102308.

Griffith, T., Rankin, K., and White-Traut, R. (2017). The Relationship Between Behavioral States and Oral Feeding Efficiency in Preterm Infants. *Advances in Neonatal Care*, 17(1):E12–E19.

Grunau, R. E., Oberlander, T., Holsti, L., and Whitfield, M. F. (1998). Bedside application of the Neonatal Facial Coding System in pain assessment of premature infants. *Pain*, 76(3):277–286.

Gui, L.-Y., Wang, Y.-X., Ramanan, D., and Moura, J. M. F. (2018). Few-Shot Human Motion Prediction via Meta-Learning. pages 432–450.

Guo, J. and Deng, J. (2019). Insightface: 2d and 3d face analysis project.

Guo, Y., Codella, N. C., Karlinsky, L., Codella, J. V., Smith, J. R., Saenko, K., Rosing, T., and Feris, R. (2020). A Broader Study of Cross-Domain Few-Shot Learning. arXiv:1912.07200 [cs].

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., and Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48. Publisher: Elsevier.

Gupta, A., Thadani, K., and O'Hare, N. (2020). Effective Few-Shot Classification with Transfer Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1061–1066, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haleem, A., Javaid, M., and Khan, I. H. (2019). Current status and applications of Artificial Intelligence (AI) in medical field: An overview. *Current Medicine Research and Practice*, 9(6):231–237.

Hashemi, J., Spina, T. V., Tepper, M., Esler, A., Morellas, V., Papanikolopoulos, N., and Sapiro, G. (2012). A computer vision approach for the assessment of autism-related behavioral markers. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–7. IEEE.

Hatfield, L. A. and Ely, E. A. (2015). Measurement of Acute Pain in Infants: A Review of Behavioral and Physiological Variables. *Biological Research For Nursing*, 17(1):100–111.

He, D., Zhou, Z., Gan, C., Li, F., Liu, X., Li, Y., Wang, L., and Wen, S. (2019). Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8401–8408. Issue: 01.

He, R., Liu, S., He, S., and Tang, K. (2022). Multi-Domain Active Learning: Literature Review and Comparative Study. *IEEE Transactions on Emerging Topics in Computational Intelligence*, pages 1–14. Conference Name: IEEE Transactions on Emerging Topics in Computational Intelligence.

Heilbron, F. C., Lee, J.-Y., Jin, H., and Ghanem, B. (2018). What do I Annotate Next? An Empirical Study of Active Learning for Action Localization. pages 199–216.

Hesse, N., Bodensteiner, C., Arens, M., Hofmann, U. G., Weinberger, R., and Sebastian Schroeder, A. (2019). Computer Vision for Medical Infant Motion Analysis: State of the Art and RGB-D Data Set. In Leal-Taixé, L. and Roth, S., editors, *Computer Vision – ECCV 2018 Workshops*, volume 11134, pages 32–49. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Hesse, N., Pujades, S., Romero, J., Black, M. J., Bodensteiner, C., Arens, M., Hofmann, U. G., Tacke, U., Hadders-Algra, M., and Weinberger, R. (2018). Learning an infant body model from RGB-D data for accurate full body motion analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 792–800. Springer.

Hesse, N., Schröder, A. S., Müller-Felber, W., Bodensteiner, C., Arens, M., and Hofmann, U. G. (2017). Body pose estimation in depth images for infant motion analysis. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1909–1912. IEEE.

Hicks, J. H. and Fairchild, K. D. (2013). Heart Rate Characteristics in the NICU: What Nurses Need to Know. *Advances in Neonatal Care*, 13(6):396.

Hodges, E. A., Johnson, S. L., Hughes, S. O., Hopkinson, J. M., Butte, N. F., and Fisher, J. O. (2013). Development of the responsiveness to child feeding cues scale. *Appetite*, 65:210–219.

Hodges, E. A., Wasser, H. M., Colgan, B. K., and Bentley, M. E. (2016). Development of Feeding Cues During Infancy and Toddlerhood. *MCN: The American Journal of Maternal/Child Nursing*, 41(4):244–251.

Holsti, L., Grunau, R. E., Oberlander, T. F., Whitfield, M. F., and Weinberg, J. (2005). Body Movements: An Important Additional Factor in Discriminating Pain From Stress in Preterm Infants. *The Clinical Journal of Pain*, 21(6):491–498.

Holub, A., Perona, P., and Burl, M. C. (2008). Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. ISSN: 2160-7508.

Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3):185–203. Publisher: Elsevier.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2022). Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5149–5169. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian Active Learning for Classification and Preference Learning. arXiv:1112.5745 [cs, stat].

Huang, X., Wan, M., Luan, L., Tunik, B., and Ostadabbas, S. (2022). Computer Vision to the Rescue: Infant Postural Symmetry Estimation from Incongruent Annotations. *arXiv preprint arXiv:2207.09352*.

Hudson-Barr, D., Capper-Michel, B., Lambert, S., Mizell Palermo, T., Morbeto, K., and Lombardo, S. (2002). Validation of the Pain Assessment in Neonates (PAIN) Scale with the Neonatal Infant Pain Scale (NIPS). *Neonatal Network*, 21(6):15–21.

Hung, H.-Y., Hsu, Y.-Y., and Chang, Y.-J. (2013). Comparison of Physiological and Behavioral Responses to Fresh and Thawed Breastmilk in Premature Infants—A Preliminary Study. *Breastfeeding Medicine*, 8(1):92–98.

Jadon, S. and Srinivasan, A. A. (2021). Improving siamese networks for one-shot learning using kernel-based activation functions. In *Data management, analytics and innovation*, pages 353–367. Springer.

Jamal, M. A. and Qi, G.-J. (2019). Task Agnostic Meta-Learning for Few-Shot Learning. pages 11719–11727.

Joshi, A. J., Porikli, F., and Papanikolopoulos, N. (2009). Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. ISSN: 1063-6919.

Kadir, T., Bowden, R., Ong, E.-J., and Zisserman, A. (2004). Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition. In *BMVC*, pages 1–10.

Kamran, F., Khatoonabadi, A. R., Aghajanzadeh, M., Ebadi, A., Faryadras, Y., and Sagheb, S. (2020). Effectiveness of Cue-Based Feeding Versus Scheduled Feeding in Preterm Infants Using Comprehensive Feeding Assessment Scales: A Randomized Clinical Trial. *Iranian Journal of Pediatrics*, 30(6).

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., and Zisserman, A. (2017). The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs].

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. arXiv:1609.04836 [cs, math].

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs].

Kirk, A. T., Alder, S. C., and King, J. D. (2007). Cue-based oral feeding clinical pathway results in earlier attainment of full oral feeding in premature infants. *Journal of Perinatology*, 27(9):572–578. Publisher: Nature Publishing Group.

Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, page 0. Lille.

Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., and Gong, B. (2021). MoViNets: Mobile Video Networks for Efficient Video Recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16015–16025, Nashville, TN, USA. IEEE.

Koonce, B. (2021). MobileNetV3. In Koonce, B., editor, *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pages 125–144. Apress, Berkeley, CA.

Kurt Sezer, H. and Küçükoğlu, S. (2020). Cue Based Feeding. *Turkiye Klinikleri Journal of Pediatrics*, 29(1):39–46.

Li, C., Pourtaherian, A., van Onzenoort, L., Ten, W. E. T. a., and de With, P. H. N. (2021). Infant Facial Expression Analysis: Towards a Real-Time Video Monitoring System Using R-CNN and HMM. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1429–1440.

Liotto, N., Cresi, F., Beghetti, I., Roggero, P., Menis, C., Corvaglia, L., Mosca, F., Aceti, A., and on behalf of the Study Group on Neonatal Nutrition and Gastroenterology—Italian Society of Neonatology (2020). Complementary Feeding in Preterm Infants: A Systematic Review. *Nutrients*, 12(6):1843. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Lorbach, M., Poppe, R., and Veltkamp, R. C. (2019). Interactive rodent behavior annotation in video using active learning. *Multimedia Tools and Applications*, 78(14):19787–19806.

Lowe, D. G. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision*, 1(1):57–72.

Lowell, D., Lipton, Z. C., and Wallace, B. C. (2018). Practical obstacles to deploying active learning. *arXiv preprint arXiv:1807.04801*.

Ludwig, S. M. and Waitzman, K. A. (2007). Changing Feeding Documentation to Reflect Infant-Driven Feeding Practice. *Newborn and Infant Nursing Reviews*, 7(3):155–160.

Lumley, J. (2003). Defining the problem: the epidemiology of preterm birth. *BJOG: An International Journal of Obstetrics & Gynaecology*, 110(s20):3–7. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1471-0528.2003.00011.x.

Maragos, P., Potamianos, A., and Gros, P., editors (2008). *Multimodal Processing and Interaction: Audio, Video, Text.* Springer US, Boston, MA.

McFadden, A., Fitzpatrick, B., Shinwell, S., Tosh, K., Donnan, P., Wallace, L. M., Johnson, E., MacGillivray, S., Gavine, A., Farre, A., and Mactier, H. (2021). Cue-based versus scheduled feeding for preterm infants transitioning from tube to oral feeding: the Cubs mixed-methods feasibility study. *Health Technology Assessment*, 25(74):1–146.

Miceli, M., Schuessler, M., and Yang, T. (2020). Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):115:1–115:25.

Mohammadi, F. G., Arabnia, H. R., and Amini, M. H. (2019). On parameter tuning in meta-learning for computer vision. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 300–305. IEEE.

Morison, S. J., Holsti, L., Grunau, R. E., Whitfield, M. F., Oberlander, T. F., Chan, H. W. P., and Williams, L. (2003). Are there developmentally distinct motor indicators of pain in preterm infants? *Early Human Development*, 72(2):131–146.

Nagy, , Földesy, P., Jánoki, I., Terbe, D., Siket, M., Szabó, M., Varga, J., and Zarándy, (2021). Continuous Camera-Based Premature-Infant Monitoring Algorithms for NICU. *Applied Sciences*, 11(16):7215. Number: 16 Publisher: Multidisciplinary Digital Publishing Institute.

Navaneeth, S., Sarath, S., Amba Nair, B., Harikrishnan, K., and Prajal, P. (2020). A Deep-Learning Approach to Find Respiratory Syndromes in Infants using Thermal Imaging. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 0498–0501, Chennai, India. IEEE.

Newland, L., L'Huillier, M. W., and Petrey, B. (2013). Implementation of Cue-Based Feeding in a Level III NICU. *Neonatal Network*, 32(2):132–137.

Nyqvist, K. (2008). Early attainment of breastfeeding competence in very preterm infants. *Acta Paediatrica*, 97(6):776–781.

Olsen, M. D., Herskind, A., Nielsen, J. B., and Paulsen, R. R. (2014). Model-based motion tracking of infants. In *European Conference on Computer Vision*, pages 673–685. Springer.

Orsi, K. C. S. C., Llaguno, N. S., Avelar, A. F. M., Tsunemi, M. H., Pedreira, M. d. L. G., Sato, M. H., and Pinheiro, E. M. (2015). Effect of reducing sensory and environmental stimuli during hospitalized premature infant sleep. *Revista da Escola de Enfermagem da USP*, 49:0550–0555. Publisher: Universidade de São Paulo, Escola de Enfermagem.

O' Mahony, N., Campbell, S., Carvalho, A., Krpalkova, L., Hernandez, G. V., Harapanahalli, S., Riordan, D., and Walsh, J. (2019). One-Shot Learning for Custom Identification Tasks; A Review. *Procedia Manufacturing*, 38:186–193.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359. Publisher: IEEE.

Park, J. (2020). Sleep Promotion for Preterm Infants in the NICU. *Nursing for Women's Health*, 24(1):24–35.

Paul, I. M., Williams, J. S., Anzman-Frasca, S., Beiler, J. S., Makova, K. D., Marini, M. E., Hess, L. B., Rzucidlo, S. E., Verdiglione, N., Mindell, J. A., and Birch, L. L. (2014). The Intervention Nurses Start Infants Growing on Healthy Trajectories (INSIGHT) study. *BMC Pediatrics*, 14(1):184.

Pereira, A. L. d. S. T., Guinsburg, R., Almeida, M. F. B. d., Monteiro, A. C., Santos, A. M. N. d., and Kopelman, B. I. (1999). Validity of behavioral and physiologic parameters for acute pain assessment of term newborn infants. *Sao Paulo Medical Journal*, 117(2):72–80.

Platt, M. J. (2014). Outcomes in preterm infants. *Public Health*, 128(5):399–403.

Puckett, B., Grover, V., Holt, T., and Sankaran, K. (2008). Cue-Based Feeding for Preterm Infants: A Prospective Trial. *American Journal of Perinatology*, 25(10):623–628.

Ravishankar, H., Sudhakar, P., Venkataramani, R., Thiruvenkadam, S., Annangi, P., Babu, N., and Vaidya, V. (2016). Understanding the mechanisms of deep transfer learning for medical images. In *Deep learning and data labeling for medical applications*, pages 188–196. Springer.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. (2021). Data Augmentation Can Improve Robustness. arXiv:2111.05328 [cs, stat].

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. (2021). A Survey of Deep Active Learning. *ACM Computing Surveys*, 54(9):180:1–180:40.

Ritu, J. T., Shakil, M. S. H., Hasan, M. N. I., Al Mamun, S., Kaiser, M. S., and Mahmud, M. (2022). Facial Detection for Neonatal Infant Pain Using Facial Geometry Features and LBP. In *Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering: TCCE 2021*, pages 509–518. Springer.

Romanov, S., Song, H., Valstar, M., Sharkey, D., Henry, C., Triguero, I., and Torres, M. T. (2021). Few-Shot Learning for Postnatal Gestational Age Estimation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ISSN: 2161-4407.

Romera-Paredes, B. and Torr, P. H. S. (2017). An Embarrassingly Simple Approach to Zero-Shot Learning. In Feris, R. S., Lampert, C., and Parikh, D., editors, *Visual Attributes*, pages 11–30. Springer International Publishing, Cham. Series Title: Advances in Computer Vision and Pattern Recognition.

Rothgänger, H. (2003). Analysis of the sounds of the child in the first year of age and a comparison to the language. *Early Human Development*, 75(1):55–69.

Rougier, S., Puissant, A., Stumpf, A., and Lachiche, N. (2016). Comparison of sampling strategies for object-based classification of urban vegetation from Very High Resolution satellite images. *International Journal of Applied Earth Observation and Geoinformation*, 51:60–73.

Saha, S., Singh, G., Sapienza, M., Torr, P. H., and Cuzzolin, F. (2016). Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*.

Settle, M. and Francis, K. (2019). Does the Infant-Driven Feeding Method Positively Impact Preterm Infant Feeding Outcomes? *Advances in Neonatal Care*, 19(1):51–55.

Settles, B. (2009). Active Learning Literature Survey. Technical Report, University of Wisconsin-Madison Department of Computer Sciences. Accepted: 2012-03-15T17:23:56Z.

Shaker, C. (2017). Infant-Guided, Co-Regulated Feeding in the Neonatal Intensive Care Unit. Part II: Interventions to Promote Neuroprotection and Safety. *Seminars in Speech and Language*, 38(02):106–115.

Shaker, C. S. (2013). Cue-Based Feeding in the NICU: Using the Infant's Communication as a Guide. *Neonatal Network*, 32(6):404–408.

Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Stenson, B. J., Tarnow-Mordi, W. O., Darlow, B. A., Juszczak, E., Askie, L., and Broadbent, R. (2013). Oxygen Saturation and Outcomes in Preterm Infants. *New England Journal of Medicine*, 368(22):2094–2104.

Stumpf, A., Lachiche, N., Malet, J.-P., Kerle, N., and Puissant, A. (2014). Active Learning in the Spatial Domain for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(5):2492–2507. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

Su, B.-H. (2014). Optimizing Nutrition in Preterm Infants. *Pediatrics & Neonatology*, 55(1):5–13.

Sun, Y., Hu, J., Wang, W., He, M., and de With, P. H. N. (2021). Camera-based discomfort detection using multi-channel attention 3D-CNN for hospitalized infants. *Quantitative Imaging in Medicine and Surgery*, 11(7):3059–3069.

Sun, Y., Shan, C., Tan, T., Long, X., Pourtaherian, A., Zinger, S., and de With, P. H. N. (2019). Video-based discomfort detection for infants. *Machine Vision and Applications*, 30(5):933–944.

Talej, M., Smith, E. R., Lauria, M. E., Chitale, R., Ferguson, K., and He, S. (2022). Responsive Feeding for Preterm or Low Birth Weight Infants: A Systematic Review and Meta-analysis. *Pediatrics*, 150(Supplement 1):e2022057092F.

Thoman, E. B. and Whitney, M. P. (1990). Behavioral states in infants: Individual differences and individual analyses. *Individual differences in infancy: Reliability, stability, prediction*, pages 113–135. Publisher: Lawrence Erlbaum Associates, Inc.

Thoyre, S., Park, J., Pados, B., and Hubbard, C. (2013). Developing a co-regulated, cue-based feeding practice: The critical role of assessment and reflection. *Journal of Neonatal Nursing*, 19(4):139–148.

Thoyre, S., Shaker, C., and Pridham, K. (2005). The Early Feeding Skills Assessment for Preterm Infants. *Neonatal Network*, 24(3):7–16.

Vahdani, E. and Tian, Y. (2021). Deep Learning-based Action Detection in Untrimmed Videos: A Survey. arXiv:2110.00111 [cs].

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Ventura, A. K. and Mennella, J. A. (2017). An Experimental Approach to Study Individual Differences in Infants' Intake and Satiation Behaviors during Bottle-Feeding. *Childhood Obesity*, 13(1):44–52.

Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., and Wierstra, D. (2016). Matching Networks for One Shot Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Vogel, J. P., Chawanpaiboon, S., Moller, A.-B., Watananirun, K., Bonet, M., and Lumbiganon, P. (2018). The global epidemiology of preterm birth. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 52:3–12.

Vrajitoru, D. (2000). Large Population or Many Generations for Genetic Algorithms? Implications in Information Retrieval. In Kacprzyk, J., Crestani, F., and Pasi, G., editors, *Soft Computing in Information Retrieval*, volume 50, pages 199–222. Physica-Verlag HD, Heidelberg. Series Title: Studies in Fuzziness and Soft Computing.

Wang, M. and Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34. Publisher: ACM New York, NY, USA.

Watson, J. and McGuire, W. (2016). Responsive versus scheduled feeding for preterm infants. *Cochrane Database of Systematic Reviews*, 2016(8).

Weinzaepfel, P., Harchaoui, Z., and Schmid, C. (2015). Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3164–3172.

Wellington, A. and Perlman, J. M. (2015). Infant-driven feeding in premature infants: a quality improvement project. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 100(6):F495–F500. Publisher: BMJ Publishing Group.

Whetten, C. H. (2016). Cue-based feeding in the NICU. *Nursing for Women's Health*, 20(5):507–510. Publisher: Elsevier.

Wilson, G. and Cook, D. J. (2020). A Survey of Unsupervised Deep Domain Adaptation. arXiv:1812.02849 [cs, stat] version: 2.

Xian, Y., Schiele, B., and Akata, Z. (2017). Zero-Shot Learning — The Good, the Bad and the Ugly. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3077–3086, Honolulu, HI. IEEE.

Xue, Y., Liao, X., Carin, L., and Krishnapuram, B. (2007). Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research*, 8(1).

Yang, C.-Y., Jiang, Z., Gu, S.-Y., Hwang, J.-N., and Yoo, J.-H. (2022). Unsupervised Domain Adaptation Learning for Hierarchical Infant Pose Recognition with Synthetic Data. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06. ISSN: 1945-788X.

Yang, Y., Ma, Z., Nie, F., Chang, X., and Hauptmann, A. G. (2015). Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization. *International Journal of Computer Vision*, 113(2):113–127.

Yang, Y., Saleemi, I., and Shah, M. (2013). Discovering Motion Primitives for Unsupervised Grouping and One-Shot Learning of Human Actions, Gestures, and Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1635–1648.

Yun, S., Oh, S. J., Heo, B., Han, D., and Kim, J. (2020). VideoMix: Rethinking Data Augmentation for Video Classification. arXiv:2012.03457 [cs].

Zaidi, Z., Rockich-Winston, N., Chow, C., Martin, P. C., Onumah, C., and Wyatt, T. (2023). Whiteness theory and the (in)visible hierarchy in medical education. *Medical Education*, n/a(n/a). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/medu.15124.

Zamzmi, G., Goldgof, D., Kasturi, R., Sun, Y., and Ashmeade, T. (2019). Machine-based Multimodal Pain Assessment Tool for Infants: A Review. arXiv:1607.00331 [cs].

Zamzmi, G., Pai, C.-Y., Goldgof, D., Kasturi, R., Ashmeade, T., and Sun, Y. (2022). A Comprehensive and Context-Sensitive Neonatal Pain Assessment Using Computer Vision. *IEEE Transactions on Affective Computing*, 13(1):28–45.

Zhang, N., Ruan, M., Wang, S., Paul, L., and Li, X. (2022). Discriminative Few Shot Learning of Facial Dynamics in Interview Videos for Autism Trait Classification. *IEEE Transactions on Affective Computing*, pages 1–1. Conference Name: IEEE Transactions on Affective Computing.

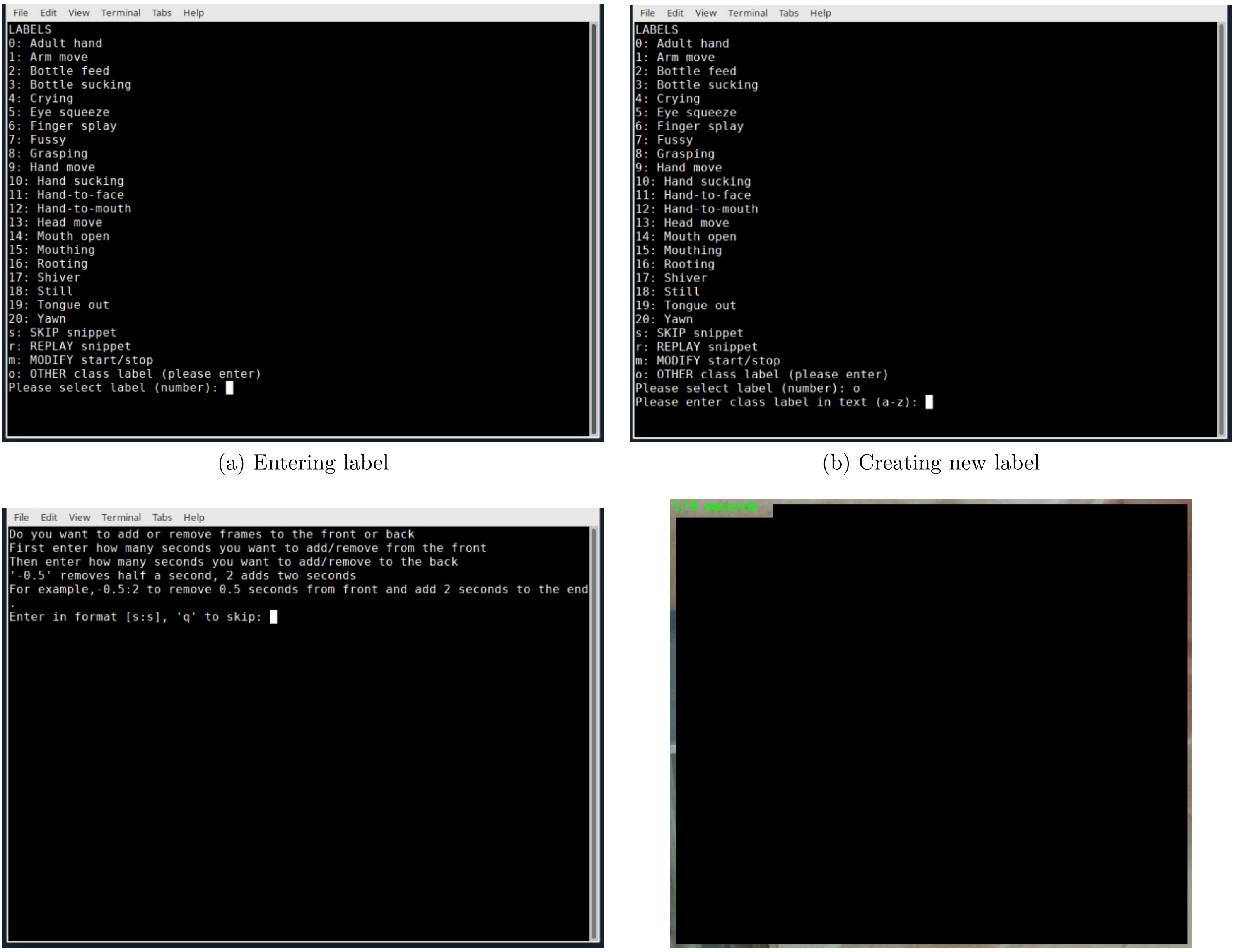
Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S., and E, W. (2021). Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning. arXiv:2010.05627 [cs, math, stat].

Zhu, Z., Zhu, X., Ye, Y., Guo, Y.-F., and Xue, X. (2011). Transfer active learning. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2169–2172, New York, NY, USA. Association for Computing Machinery.

Zieren, J. and Kraiss, K.-F. (2005). Robust person-independent visual sign language recognition. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 520–528. Springer.

Zimmerman, E. (2013). Limited evidence suggests that *ad libitum* or demand/semi-demand feeding allows for earlier hospital discharge compared to scheduled feeds for preterm infants. *Evidence-Based Communication Assessment and Intervention*, 7(2):63–67.

A Experimental Design



```
LABELS
0: Adult hand
1: Arm move
2: Bottle feed
3: Bottle sucking
4: Crying
5: Eye squeeze
6: Finger splay
7: Fussy
8: Grasping
9: Hand move
10: Hand sucking
11: Hand-to-face
12: Hand-to-mouth
13: Head move
14: Mouth open
15: Mouthing
16: Rooting
17: Shiver
18: Still
19: Tongue out
20: Yawn
s: SKIP snippet
r: REPLAY snippet
m: MODIFY start/stop
o: OTHER class label (please enter)
Please select label (number):
```

(a) Entering label

```
LABELS
0: Adult hand
1: Arm move
2: Bottle feed
3: Bottle sucking
4: Crying
5: Eye squeeze
6: Finger splay
7: Fussy
8: Grasping
9: Hand move
10: Hand sucking
11: Hand-to-face
12: Hand-to-mouth
13: Head move
14: Mouth open
15: Mouthing
16: Rooting
17: Shiver
18: Still
19: Tongue out
20: Yawn
s: SKIP snippet
r: REPLAY snippet
m: MODIFY start/stop
o: OTHER class label (please enter)
Please select label (number): o
Please enter class label in text (a-z):
```

(b) Creating new label

```
Do you want to add or remove frames to the front or back
First enter how many seconds you want to add/remove from the front
Then enter how many seconds you want to add/remove to the back
'-0.5' removes half a second, 2 adds two seconds
For example,-0.5:2 to remove 0.5 seconds from front and add 2 seconds to the end
.
Enter in format [s:s], 'q' to skip:
```

(c) Modifying snippet

[redacted]

(d) Snippet shown as video

Figure A.1: User interface for AL user study, showing the command line interface for (a) entering a label, (b) entering a label that is not available in the SLAPI dataset and (c) the instructions for modifying a snippet. (d) shows a blacked-out example a a snippet being played. Only a subset of the classes are shown.

Algorithm 1 Genetic Algorithm for Hyperparameter Tuning

```
**Require:** search space $S$, number of generations $G$, population size $P$, crossover probability $p_{cx}$, mutation probability $p_{mut}$, tournament size $t$
Initialize a population *pop* of size $P$ with random individuals from search space $S$
Evaluate the fitness of each individual in *pop*
**for** $g \leftarrow 1$ to $G$ **do**
    *parents* ← Select $P$ individuals from *pop* using tournament selection with tournament size $t$
    $offspring$ ← Crossover *parents* with probability $p_{cx}$ using two-point crossover
    Mutate $offspring$ with probability $p_{mut}$ using uniform mutation within the search space bounds
    Evaluate the fitness of each individual in $offspring$
    *pop* ← $offspring$
**end for**
*best* ← the individual with the best fitness in the final population *pop*
**return** *best*
```

Algorithm 2 Weighted Sample Selection for Each Class by Probability

```
**Require:** model, videos, classes, classWeights, numSamples
unlabeledProbs ← softmax(model(videos))
normalizedWeights ← normalize(classWeights)
selectedIndices ← {}
**while** length(selectedIndices) < numSamples **do**
    topIndices ← {}
    **for** classIdx in classes **do**
        classProbs ← unlabeledProbs[classIdx]
        samplesPerClass ← int(normalizedWeights[classIdx] * numSamples)
        samplesPerClass ← max(1, samplesPerClass)
        **if** length(classProbs) < samplesPerClass **then**
            Break loop
        **end if**
        topKIndices ← topK(classProbs, k=samplesPerClass)
        topIndices[classIdx] ← topKIndices
    **end for**
    Append unique(topIndices) to selectedIndices
    Update unlabeledProbs to remove indices in selectedIndices
**end while**
**return** selectedIndices
```

Algorithm 3 Proposal Generation from Untrimmed Video

```
Require: startidx, video
Set video frame index: startIdx
result $\leftarrow$ processFrames(video[startIdx : startIdx+3s])
nextResult $\leftarrow$ processFrames(video[startIdx+3s : startIdx+6s])
label, nextLabel $\leftarrow$ getLabel(result), getLabel(nextResult)
while label = nextLabel and clipLength $\leq$ 15 seconds do
    stopIdx $\leftarrow$ video frame index
    result $\leftarrow$ combine(result, nextResult)
    nextResult $\leftarrow$ processFrames(video[stopIdx : stopIdx+3s])
    label $\leftarrow$ nextLabel
    nextLabel $\leftarrow$ getLabel(nextResult)
end while
return result, startIdx, stopIdx
```

B Figures Main Experiments

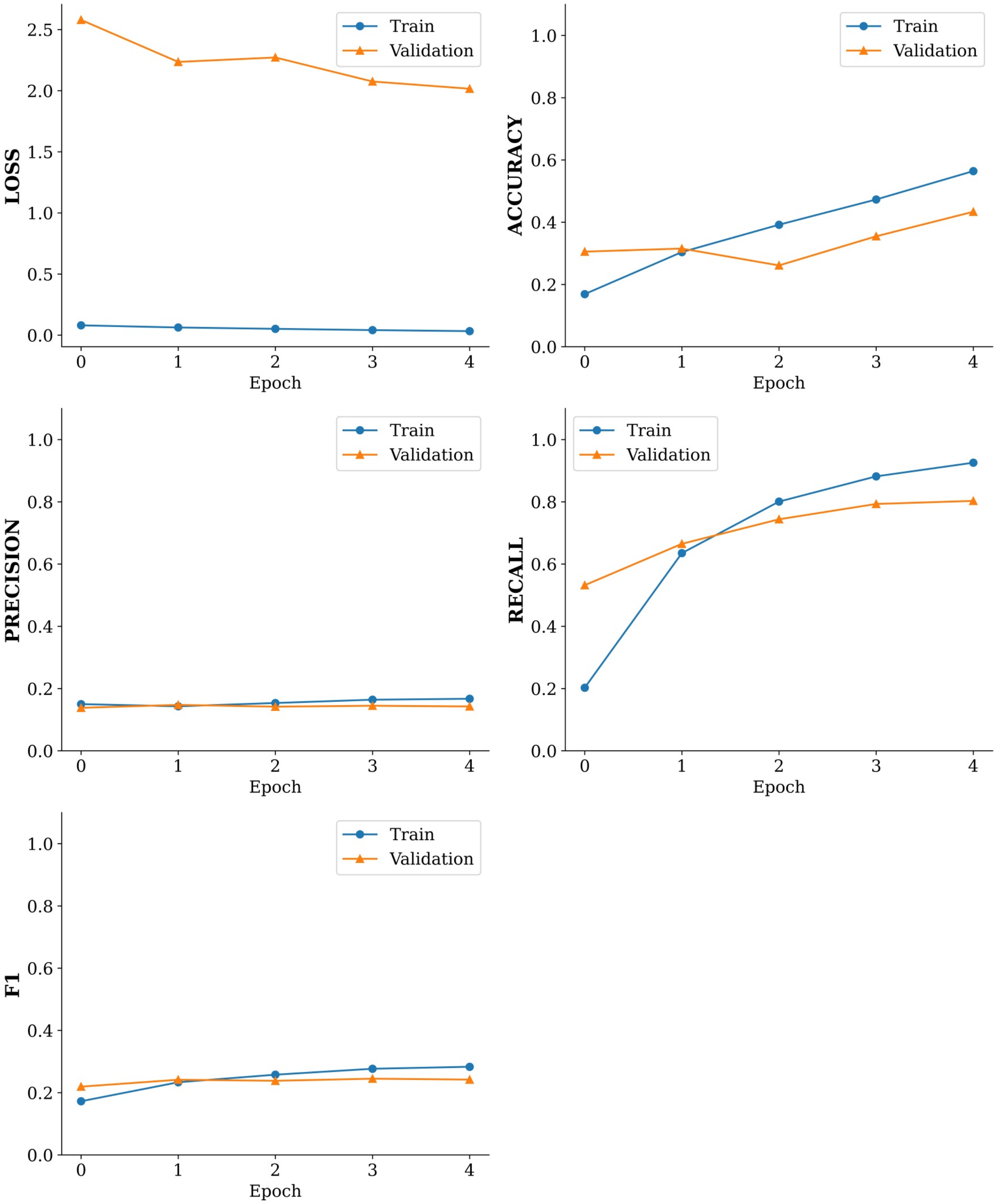


Figure B.1: Training history of the baseline first experimental trial with MoViNet A2 across different metrics. The subplots display the evolution of (top left) loss, (top right) accuracy, (middle left) precision, (middle right) recall, and (bottom) F_1 score during training.

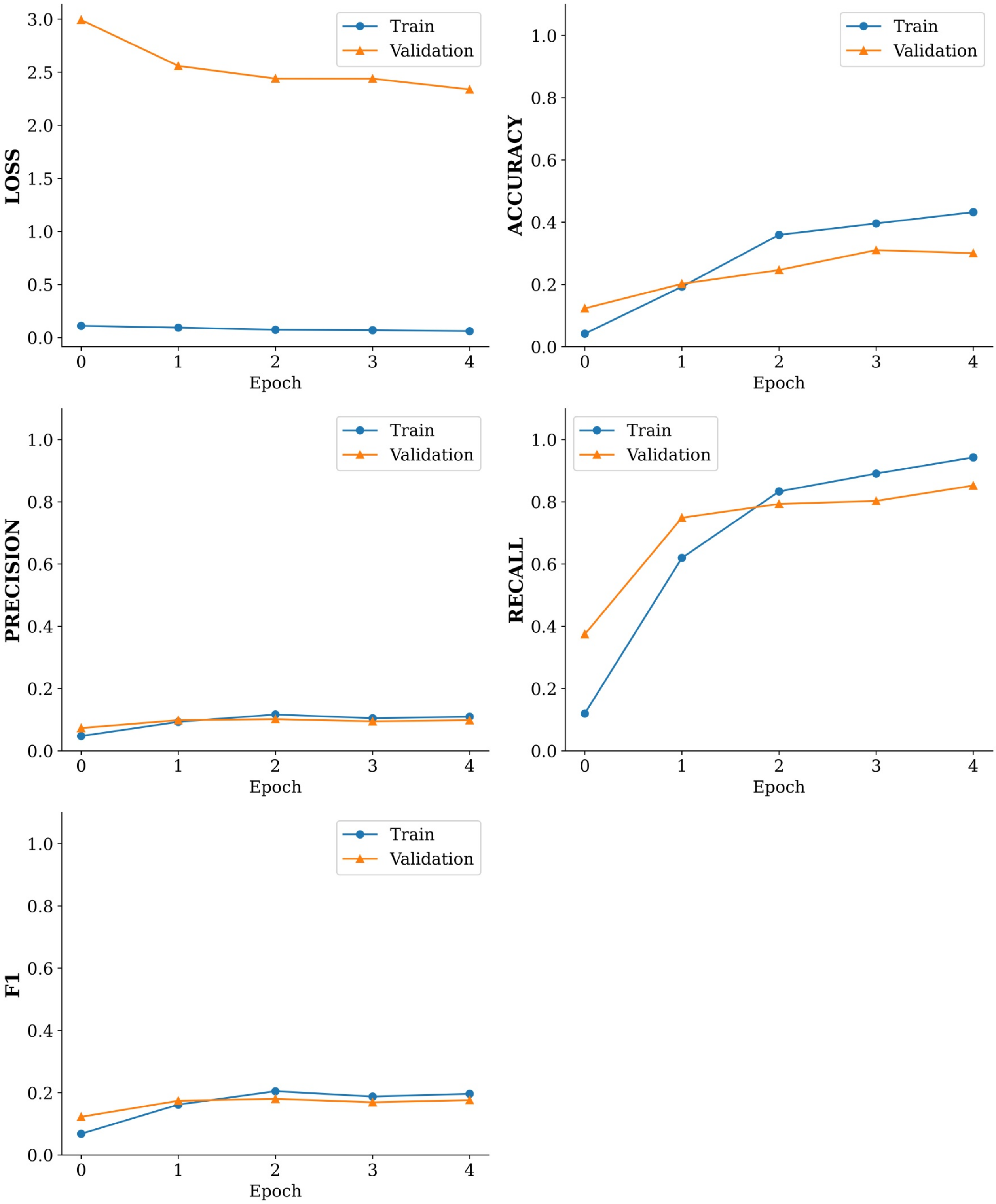


Figure B.2: Training history of the FSL first experimental trial with MoViNet A2 across different metrics. The subplots display the evolution of (top left) loss, (top right) accuracy, (middle left) precision, (middle right) recall, and (bottom) F_1 score during training.

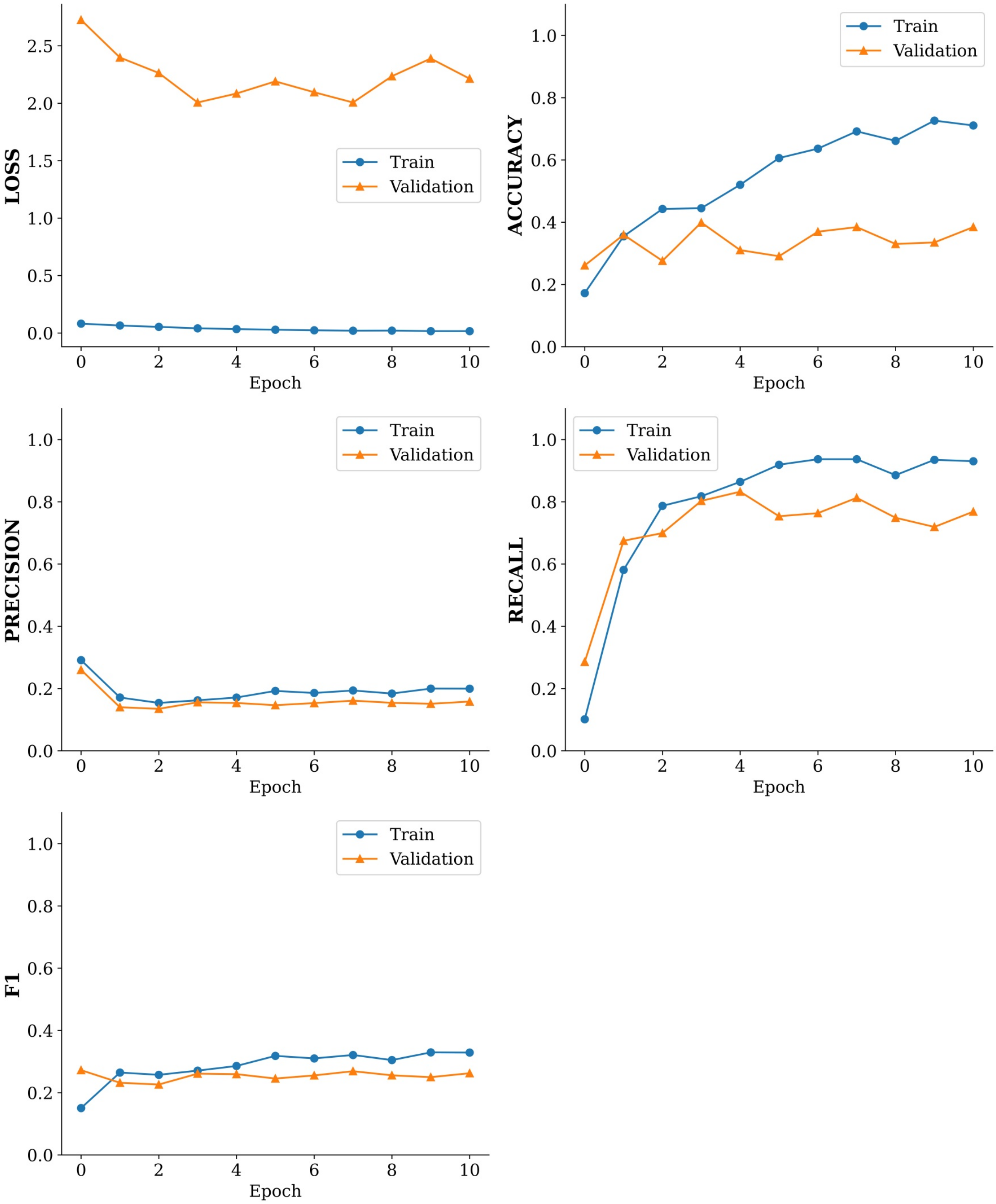


Figure B.3: Training history of the AL experiment with MoViNet A2 across different metrics. The subplots display the evolution of (top left) loss, (top right) accuracy, (middle left) precision, (middle right) recall, and (bottom) F_1 score during training. Each block of three epochs shows consecutive AL iterations.

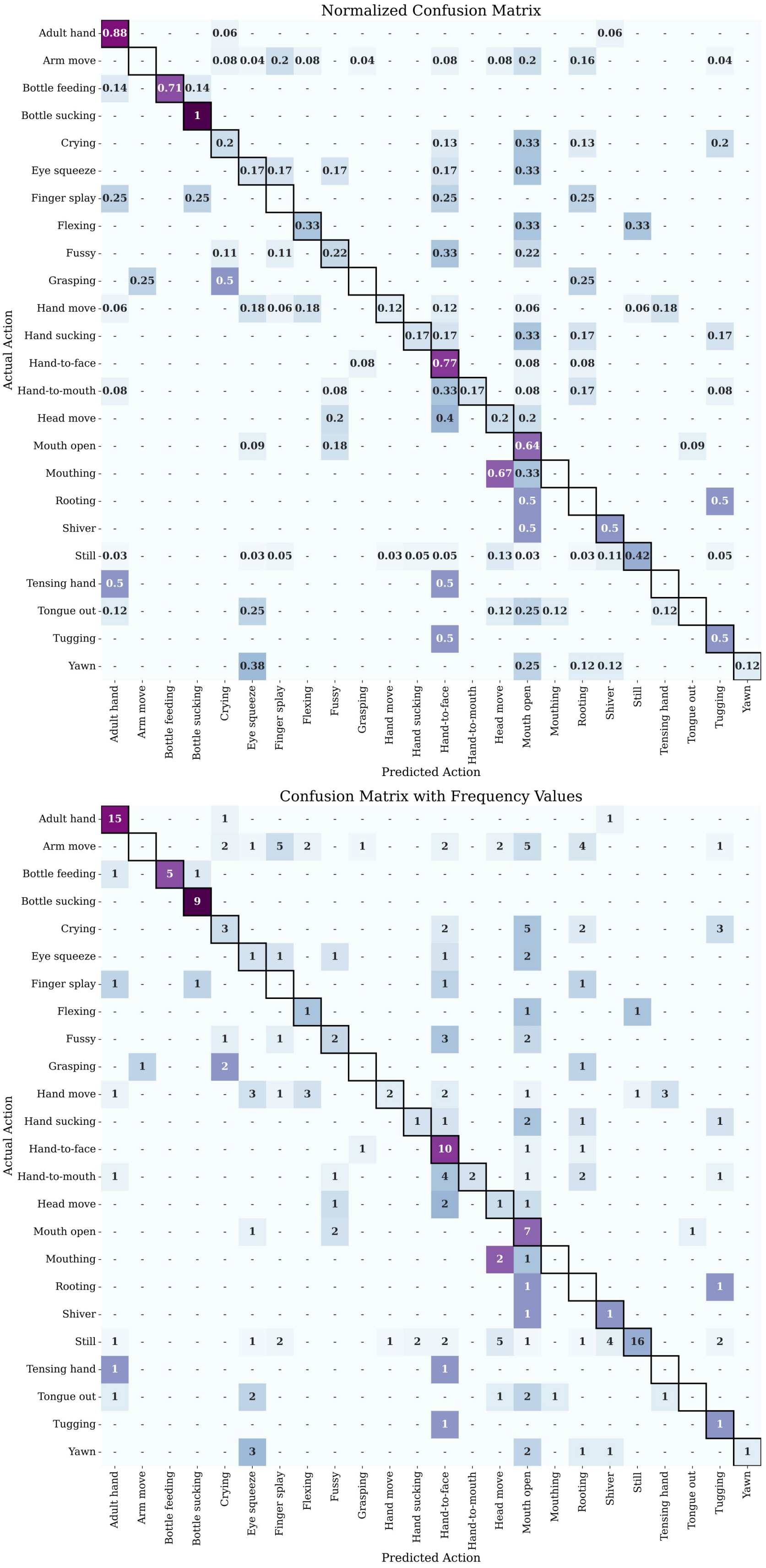
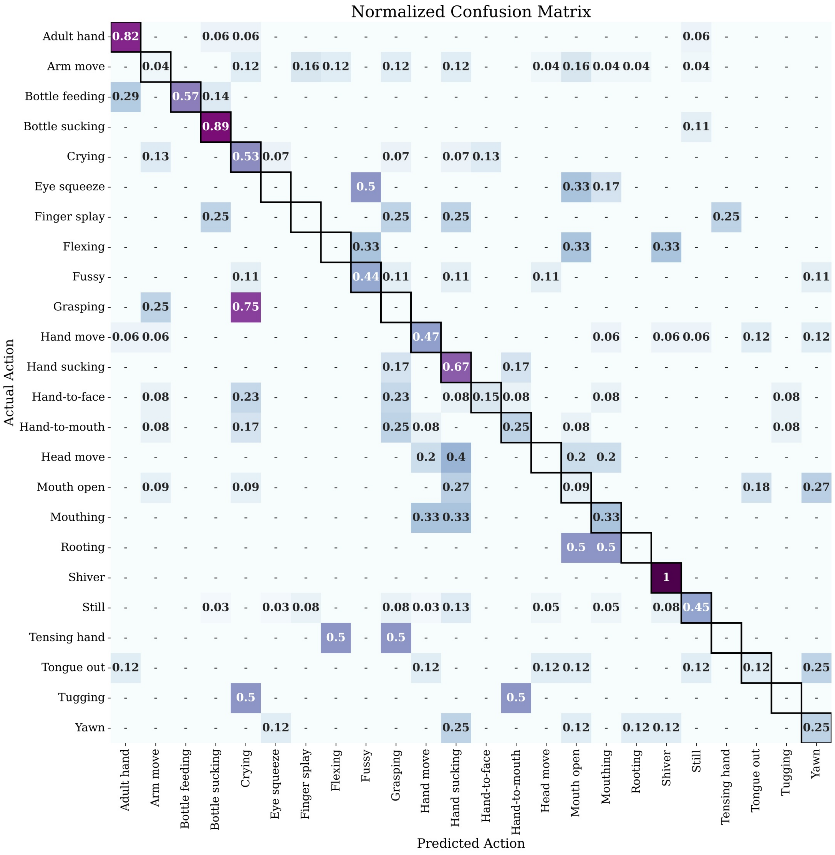


Figure B.4: Confusion matrices for FSL first experimental trial with MoViNet A2. On top, a normalized confusion matrix displaying the proportion of correct and incorrect predictions for each class. On the bottom, a frequency-based confusion matrix showing the absolute number of predictions for each class.



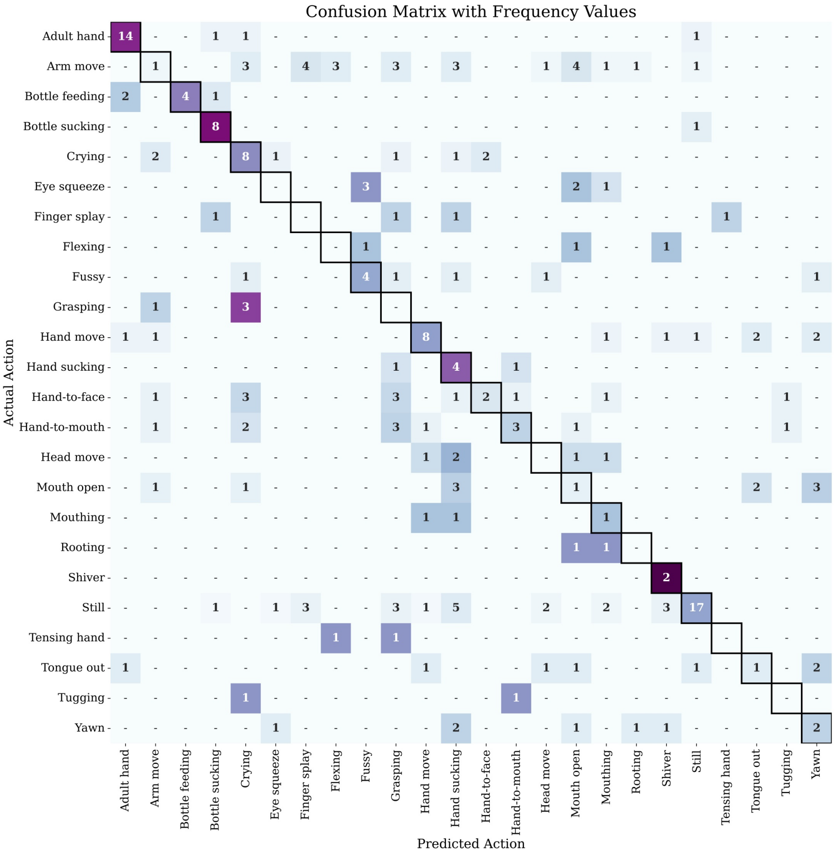
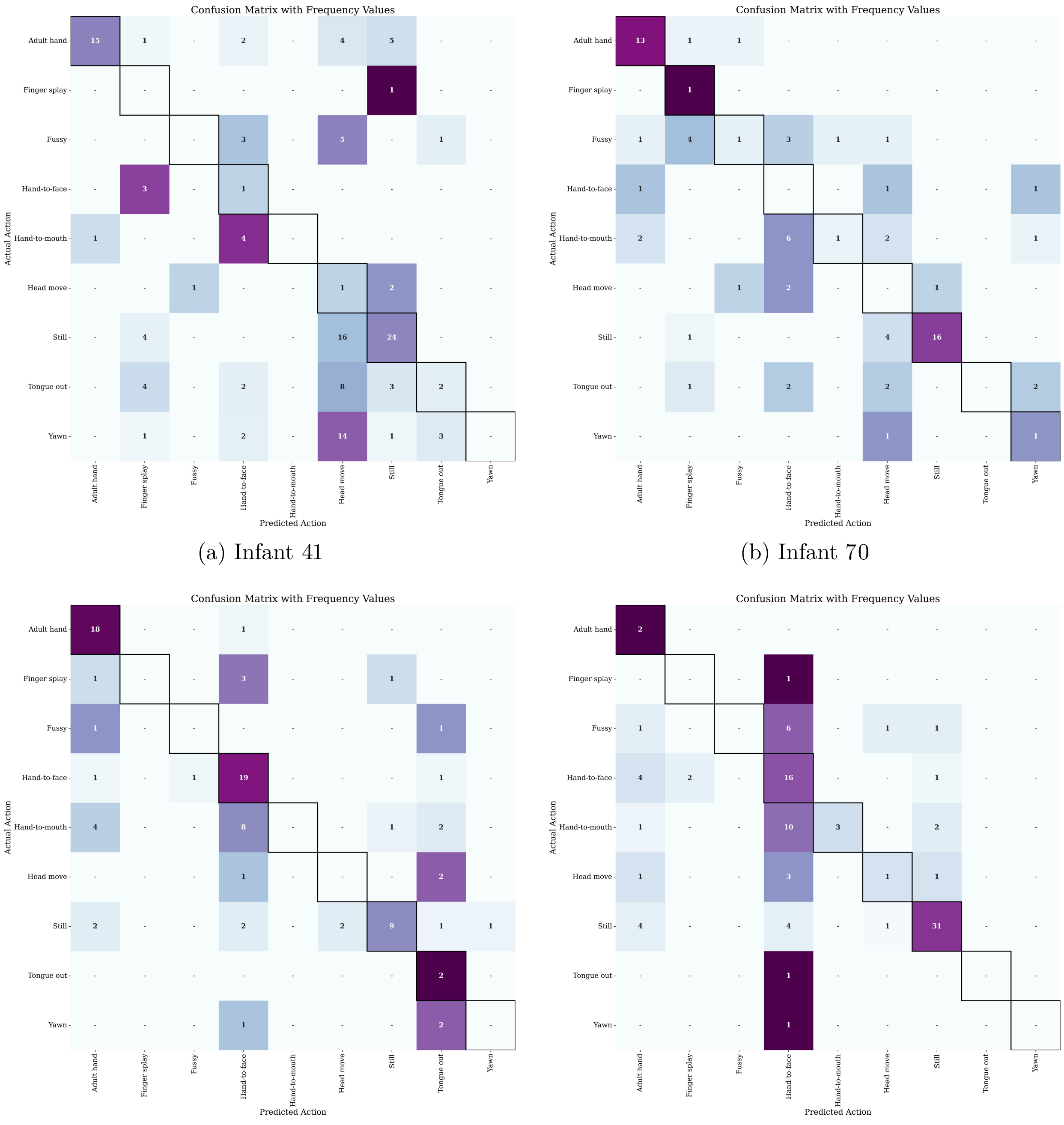


Figure B.5: Confusion matrices for AL experiment with MoViNet A2. On top, a normalized confusion matrix displaying the proportion of correct and incorrect predictions for each class. On the bottom, a frequency-based confusion matrix showing the absolute number of predictions for each class.

C Figures Ablation Study



(a) Infant 41

(b) Infant 70

(c) Infant 86

(d) Infant 61

Figure C.1: Confusion matrices representing the prediction results for four different infants in the test set under ablation condition V3. Each matrix illustrates the predicted versus actual classes, providing insights into the model's performance specific to each infant.

Class	Iteration 1	Iteration 2	Iteration 3	Total
Arm move	7	4	12	23
Crying	—	—	4	4
Eye squeeze	—	1	3	4
Finger splay	1	1	2	4
Fussy	2	—	1	3
Grasping	—	—	2	2
Hand move	3	1	—	4
Hand-to-face	2	1	2	5
Hand-to-mouth	—	—	3	3
Head move	3	—	12	15
Shiver	2	1	1	4
Still	29	41	5	75
Tugging	—	—	2	2
Yawn	—	—	1	1
Other	1	—	—	1

Table C.1: V12 — The distribution of classes identified during each iteration of the AL experiment. The table highlights the frequency of each class per iteration and provides a total sum of instances across all iterations. The class "*Still*" was the most frequently identified across the iterations, with a total of 75 instances, while "*Arm move*" and "*Head move*" were the next most common, with 23 and 15 instances respectively. Less frequent classes such as "*Crying*', "*Eye squeeze*", "*Finger splay*", and "*Yawn*" also emerged in the latter stages of the experiment