MASTER's degree Thesis
Double MASTER's degree in MECHANICAL & MECHATRONICS ENGINEERING

VIDEO-BASED AUTOMATIC MONITORING OF PATIENTS IN THE INTERMEDIATE CARE UNIT

Supervisors

Prof. Teresa BERRUTI
Prof. Massimiliano COLARIETI-TOSTI

Candidate

Chiara Noémie LLINAS

October 2025

Acknowledgements

I would like to thank Katarina, head of the nursing staff, for her unwavering support, involvement, and trust. Her enthusiasm and leadership were a constant source of motivation and played a decisive role in driving this project forward.
I am also sincerely thankful to all the nurses and assistant nurses who contributed their time and energy, especially Maria and Catalina from the hospital in Solna, and Sami and Kristian from the hospital in Huddinge, for their engagement and collaboration throughout the project.
Many thanks as well to Amélie H. and Professor Jean D.V., whose encouragement and support have been greatly appreciated.
I would also like to thank Professor Teresa Berruti, my academic advisor at Politecnico di Torino, for her guidance, and my teammates Emeryne, Quentin, and Aymeric for their hard work, commitment, and team spirit throughout this challenging and rewarding experience.
Finally, I wish to express my deepest gratitude to Professor Massimiliano Colarieti-Tosti, Mamo, my supervisor at KTH Royal Institute of Technology, for his kindness, valuable guidance, and trust over these six months.

Abstract (English)

This thesis investigates the development of an automatic monitoring system for patients in Intermediate Care Units (IMUs) at Karolinska University Hospital in Solna and Huddinge, Stockholm. The project aims to support nurses by reducing the burden of continuous visual supervision and ensuring the detection of patient agitation or risk-related behaviors. Two deep learning architectures were explored and compared: MoViNet-A5, a lightweight 3D convolutional neural network optimized for real-time video recognition, and PredFormer, a transformer-based model designed for capturing long-range spatiotemporal dependencies. The models were fine-tuned on a custom dataset, initially composed of simulation videos collected with a multi-camera Raspberry Pi setup, with the intention of later extending to real patient data under ethical approval.

In addition to video recognition, preliminary work on multimodal integration of video, electrocardiogram (ECG), and audio signals was conducted through attention-based mechanisms, highlighting the potential benefits of combining heterogeneous data sources. A strong collaboration with nursing staff played a central role, ensuring that the system was aligned with clinical workflows and ethical requirements.

The results demonstrate the feasibility of deploying lightweight models like MoViNet for real-time monitoring in hospital rooms. Although limited by the lack of real patient data during this thesis, the simulations validated the technical pipeline and prepared the ground for future research. The findings suggest that deep learning-based monitoring systems could significantly improve patient safety and reduce nurse workload, provided that ethical, technical, and clinical challenges are carefully addressed.

Abstract (Svenska)

Denna avhandling undersöker utvecklingen av ett automatiskt övervakningssystem för patienter på intermediärvårdsavdelningar (IMU) vid Karolinska Universitetssjukhuset i Solna och Huddinge, Stockholm. Projektet syftar till att stödja sjuksköterskor genom att minska bördan av kontinuerlig visuell övervakning och säkerställa upptäckt av patienters agitation eller riskrelaterade beteenden. Två djupinlärningsarkitekturer undersöktes och jämfördes: MoViNet-A5, ett lättviktigt 3D-konvolutionsneuronätverk optimerat för videigenkänning i realtid, och PredFormer, en transformatorbaserad modell utformad för att fånga långväga rumsliga och tidsmässiga beroenden. Modellerna finjusterades på en anpassad dataset, som initialt bestod av simuleringsvideor insamlade med en Raspberry Pi-uppsättning med flera kameror, med avsikten att senare utvidga till verkliga patientdata under etiskt godkännande.

Utöver videigenkänning genomfördes preliminärt arbete med multimodal integration av video, elektrokardiogram (EKG) och ljudsignaler genom uppmärksamhetsbaserade mekanismer, vilket belyste de potentiella fördelarna med att kombinera heterogena datakällor. Ett starkt samarbete med vårdpersonal spelade en central roll för att säkerställa att systemet var anpassat till kliniska arbetsflöden och etiska krav.

Resultaten visar att det är möjligt att använda lätta modeller som MoViNet för realtidsövervakning på sjukhusrum. Även om simuleringarna begränsades av bristen på verkliga patientdata under arbetet med denna avhandling, validerade de den tekniska processen och banade väg för framtida forskning. Resultaten tyder på att övervakningssystem baserade på djupinlärning skulle kunna förbättra patientsäkerheten avsevärt och minska sjuksköterskornas arbetsbelastning, förutsatt att etiska, tekniska och kliniska utmaningar hanteras noggrant.

Table of Contents

List of Figures……..10

List of Tables……11

List of Acronyms……12

1. Introduction……15
   1.1. Clinical context and problem statement……..15
   1.2. Motivation and relevance of the study……16
   1.3. Objectives of the thesis……...18
   1.4. Thesis structure……...19
   1.5. Working environment……19
      1.5.1. Karolinska hospitals……...19
      1.5.2. Patients admitted in the IMU……20
      1.5.3. Team organization……..23

2. State of the Art……25
   2.1. Deep learning..……...…….25
      2.1.1. Foundations and applications……...25
      2.1.2.Deep learning for action recognition…….……...26
      2.1.3. CNNs, RNNs, and transformers……...27
   2.2. Video-based patient monitoring…….....29
   2.3. Multimodal monitoring……31
   2.4. Ethical and privacy considerations in healthcare AI……32
   2.5. Summary research gaps……33

3. Materials and methods……..35
   3.1. Ethical approval and patient consent……...35
   3.2. Data collection setup……..36
      3.2.1. Camera system and raspberry Pi setup…….……..36
      3.2.2. Control box and labeling mechanism……..39
   3.3. Dataset construction and preprocessing……..40
      3.3.1. Simulation data collection…….....40
      3.3.2. Preprocessing pipeline and data augmentation……42
      3.3.3. Guided sampling strategy……43
   3.4. Model architectures……....44
      3.4.1. MoViNet……...44

3.4.2. PredFormer……..47
3.4.3. Preliminary comparison……49
3.4.4. Multimodal integration and attention mechanism……50
3.5. Model training……51
3.5.1. Fine tuning and hyperparameters……51
3.5.2. Optimizers, loss functions, and dropout……53
3.5.3. Frameworks: Tensorflow vs Pytorch……54
3.6. Evaluation……55
3.6.1. Evaluation metrics……..55
3.6.2. Cross-validation and robustness checks……..57
3.7. Collaboration with the medical staff……58

4. Results and discussion……..61
4.1. Model training results……62
4.1.1. MoViNet performance……...62
4.1.2. PredFormer performance……...63
4.1.3. Comparative analysis……65
4.2. Discussion……66
4.2.1. Technical constraints……..66
4.2.2. Practical challenges……67
4.2.3. Ethical and clinical implications……...68

5. Conclusions and future work……...69
5.1. Summary of the work……69
5.2. Key findings……69
5.3. Limitations of the study……70
5.4. Suggestions for future research……...71

References……73

Appendixes……..78

List of Figures

Figure 1 - Approximated intervention frequency per hour depending on the patient's level of agitation……...16
Figure 2 - Average time spent in patient rooms depending on agitation level……..17
Figure 3 - Nurses' reported stress and exhaustion after shifts…….18
Figure 4 - Map of Stockholm…….20
Figure 5 - Scheme of an intracranial aneurysm……..21
Figure 6 - Scheme of the coiling and clipping technique for aneurysm treatment……...20
Figure 7 - Aneurysm coiling - case illustration……...22
Figure 8 - Cameras and Raspberry Pi setup……37
Figure 9 - Camera and button box in the hospital environment……..38
Figure 10 - Patient room equipped with the camera setup in Karolinska, Solna……..40
Figure 11 - Examples of frames associated with a tag '1' (intervention required).....41
Figure 12 - Examples of frames associated with a tag '0' (no intervention required).....41
Figure 13 - MoViNet-A5 architecture……...46
Figure 14 - Question box in the IMU of Karolinska, Solna……...58
Figure 15 - Training vs validation curves for K-fold cross-validation loss with MoViNet…..61
Figure 16 - Confusion matrix of the fine-tuned MoViNet-A5 model using the simulation dataset…….62
Figure 17 - Training vs validation curves for K-fold cross-validation loss with PredFormer……..63
Figure 18 - Confusion matrix of the fine-tuned PredFormer model using the simulation dataset…….……..64

List of Tables

Table 1 - Simulated dataset distribution……..……43
Table 2 - Preliminary comparison between MoViNet and PredFormer…….49
Table 3 - Results of the fine-tuned MoViNet-A5 model using the simulation dataset……..…62
Table 4 - Results of the fine-tuned PredFormer model using the simulation dataset……...…64

List of Acronyms

IMU: Intermediate Care Unit

ICU: Intensive Care Unit

ECG: Electrocardiogram

AI: Artificial Intelligence

PMS: Patient Monitoring System

CNN: Convolutional Neural Network

RNN: Recurrent Neural Network

GDPR: General Data Protection Regulation

FPS: Frames per Second

ANN: Artificial Neural Network

FLOP: Floating-point Operation

ViT: Vision Transformer

1. Introduction

1.1. Clinical context and problem statement

Continuous monitoring of hospitalized patients is essential to ensure patient safety and detect critical situations early. This is especially the case in IMUs, intermediary units between intensive care units (ICUs) and general hospitalization. IMUs welcome patients who need a significant level of monitoring and more frequent attention from the medical team than standard wards, but who do not require the same level of medical support as those in intensive care. In such units, health professionals must remain vigilant, as patients may present sudden changes in condition that require immediate intervention.

Today, this observation is predominantly done manually and represents a large work-load for nurses and assistant nurses. Due to the high demand on personnel work and risk of delayed response, increasing attention is being paid to the development of automatic systems to assist clinical staff. This need can be filled in this project by constructing a video-based automatic monitoring system which will make use of AI and deep learning technologies. The system is intended to detect patient behaviors in real time, and to issue alerts when staff intervention may be required. Patients admitted to intermediate care generally suffer from conditions that cause episodes of delirium. It is difficult to give an exact definition, but certain signs are characteristic. During these episodes, patients may exhibit sudden, erratic, or violent movements that can pose a direct risk to their safety. Such behavior often leads to situations where patients attempt to disconnect catheters, intravenous lines, or other essential medical equipment, which can cause severe medical complications. In other cases, patients may strike surrounding objects, collide with the bed frame, or even fall out of bed, resulting in physical injury and additional suffering.

These behaviours make working in this unit particularly challenging, since constant monitoring of patients is essential in order to instantly react, as delays can significantly increase both the physical and psychological harm experienced by the patient. By assisting medical personnel in identifying situations that require attention, such a system could improve patient outcomes, reduce staff workload, and enhance the overall quality of care in IMUs.

1.2. Motivation and relevance of the study

The starting point of this thesis lies in the work of Benjamin Jefford-Baker, whose master's project [1] provided the foundation for the current study. By leveraging transfer learning on the MoViNet-A2 architecture, and training on a dataset constructed from video extracts of medical TV shows that simulated patient behaviors, he demonstrated the feasibility of using deep learning for automatic monitoring in hospital care. Building upon this milestone and under the scope of an ethical approval, the present thesis intends to further develop and refine the monitoring system by fine-tuning a more advanced model, MoViNet-A5, and extending it toward a multimodal framework that integrates not only video, but also sound and electrocardiogram (ECG) signals.

The use of such a tool is relevant in the daily reality of the IMUs. The units involved in the project are made up of teams of around 10 to 15 nurses and assistant nurses, who generally work in pairs during shifts. A single nurse is typically responsible for two patients, continuously monitoring their vital signs on a screen while maintaining direct visual contact with the patient in the room. Interventions are frequent, ranging from once per hour for calm patients to several times per minute for agitated ones.

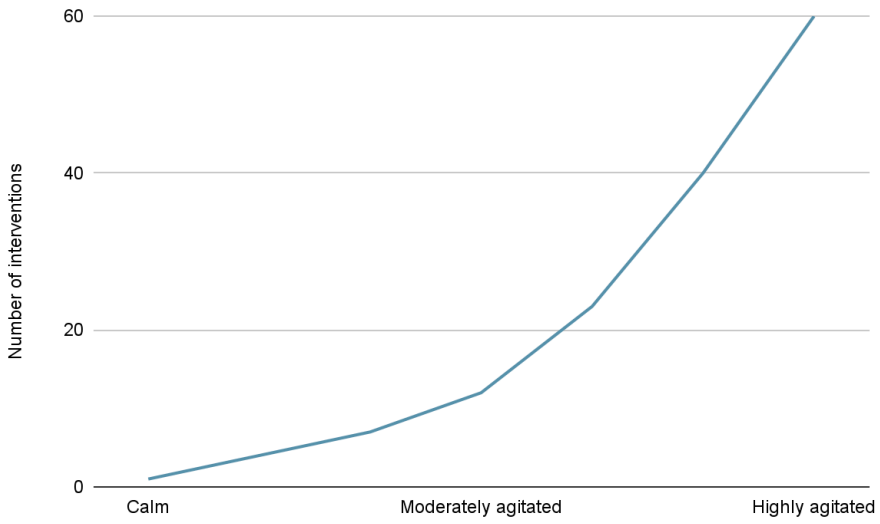


Figure 1 - Approximated intervention frequency per hour depending on the patient's level of agitation

Maintaining concentration, especially during night shifts, is particularly difficult, often leading to headaches, fatigue, and stress. Moreover, depending on the patient's level of agitation, nurses may spend between 10 and 47 minutes per hour walking to and from patient rooms.

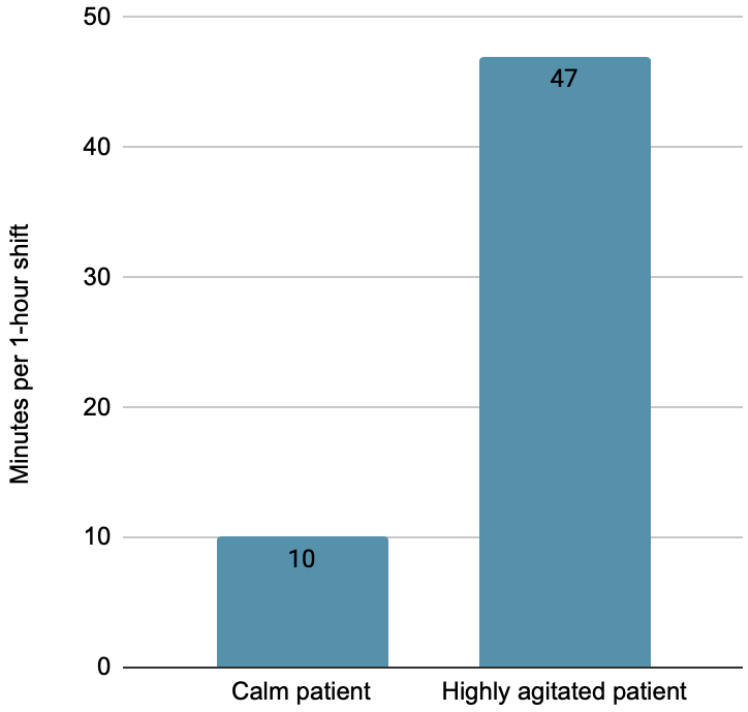


Figure 2 - Average time spent in patient rooms depending on agitation level

Therefore, this constant vigilance is both physically and emotionally tiring: according to staff, 89% report feeling emotionally stressed and 63% physically exhausted at the end of their shifts.

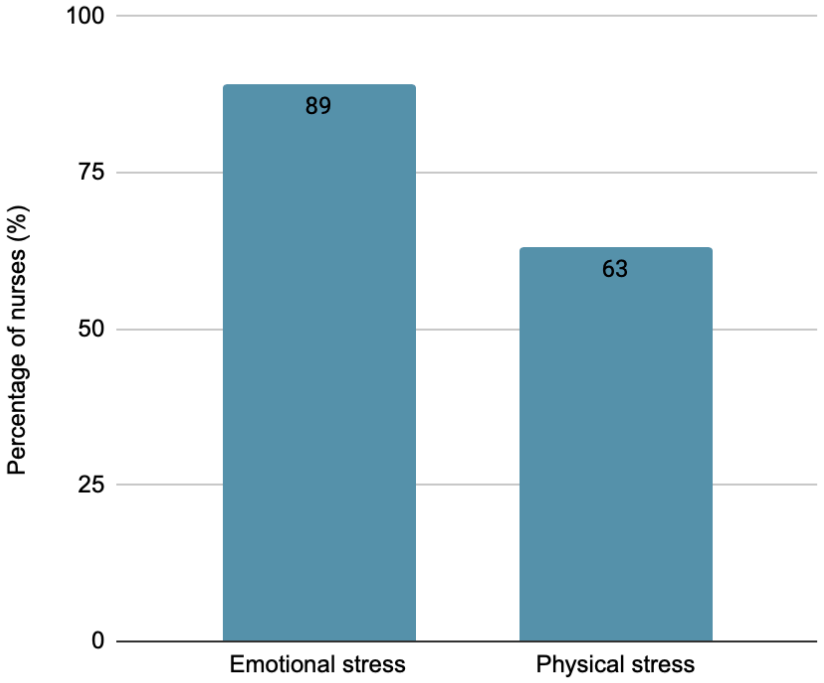


Figure 3 - Nurses' reported stress and exhaustion after shifts

1.3. Objectives of the Thesis

The objective of this thesis is twofold. First, it aims to explore and compare the performance of two state-of-the-art architectures for video action recognition: MoViNet, a lightweight convolutional model optimized for real-time inference, and PredFormer, a transformer-based model designed for accurate spatiotemporal sequence modeling. By comparing the performance of these two models, the aim is to determine their robustness, and respective ability to meet the needs of the intermediate care units. Furthermore, the addition of data channels such as sound and ECG signals and the ability of multimodal integration to enhance the model's viability is explored, although secondary to video recognition.

Beyond technical exploration, the human aspect is a central element of this thesis. By demonstrating the feasibility of a strong collaboration between technical and medical teams, this thesis not only seeks to ensure the technical compliance of the models in the reality of a hospital working environment, but also to make a meaningful contribution to improving patient safety and supporting healthcare professionals in their work.

1.4. Thesis structure

This thesis is organized into five main chapters.
The first one provides the clinical context of the project within the IMUs of Karolinska. It introduces the motivation behind developing an AI-based tool and its ethical delimitations. The state of the art is detailed in a second chapter, situating this work within the broader landscape of AI in healthcare. In a third chapter, the materials and methods used are being described. It presents the methodological aspects of the project, the architectures under investigation, the preprocessing strategies, the chosen pipeline, the setup for data collection and simulation, the technical infrastructure built on Raspberry Pi devices and multi-camera installations. It also highlights the collaboration with the nursing staff and the team organization that supported the project. Then, a fourth chapter is dedicated to the results and the discussion. It reports the results obtained with the simulation-based dataset and the fine-tuning experiments and emphasizes both the technical and practical challenges encountered during the project, the constraints related to the deployment of an AI in real-world hospital settings, and the ethical implications of our patient monitoring system. Finally, the last chapter summarizes the work and the contributions of this thesis. It highlights the key findings, and discusses both the limitations of the current study, and some directions for future research.

1.5. Working environment

1.5.1. Karolinska University Hospitals

This project was carried out in collaboration between KTH Royal Institute of Technology and the Intermediate Care Units (IMUs) of Karolinska University Hospital, located in Stockholm, Sweden. The latter is among the largest and most prestigious teaching hospitals in Europe, affiliated with Karolinska Institutet, the medical university known worldwide for awarding the Nobel Prize in Physiology or Medicine. The hospital's two main campuses are located in Solna and Huddinge, with approximately 1,600 patient beds and 15,800 employees in 2023 [2]. This strong influence of innovation and research in the medical world makes it an ideal field for the development, testing and implementation of an artificial intelligence tool in the real-world healthcare context, such as the one developed in this project.

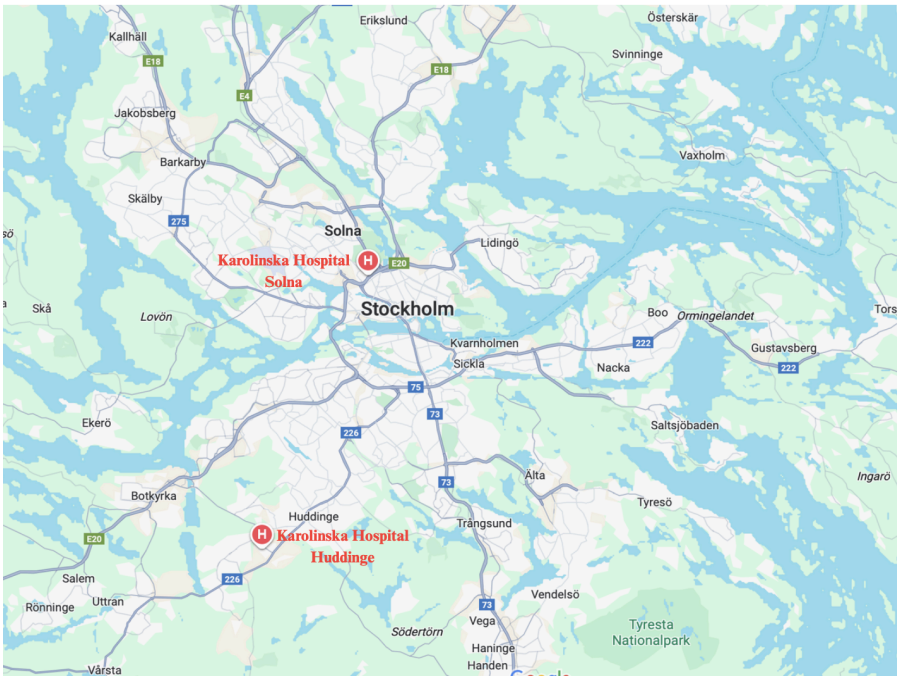


Figure 4 - Map of Stockholm

IMUs are equipped with patient vital signs monitoring equipment, Patient Monitoring Systems (PMS), and medical staff work in shifts around the clock to monitor patients. These wards welcome patients who require special attention, but do not require the level of monitoring provided in intensive care. They typically stay for 24 to 48 hours before being transferred to other wards with lower levels of supervision. The IMUs host a diverse group of patients recovering from serious conditions, such as neurological surgeries or acute medical episodes, many of whom may experience confusion, agitation, or other behavioral and cognitive changes that increase their risk of falls or sudden deterioration.

1.5.2. Patients admitted in the IMU

The patients admitted in the Intermediate Care Unit are generally suffering from neurological conditions affecting the head, neck, and spine. This includes patients recovering from brain or spinal surgeries, those with brain trauma, and individuals with vascular abnormalities.

The main pathology of individuals admitted to intermediate care in Karolinska is intracranial aneurysm. They correspond to localized dilations of blood vessels in the brain caused by a weakening of the arterial wall.

Figure 5 - Scheme of an intracranial aneurysm [46]

These aneurysms may remain asymptomatic for years, but in case of a rupture, they can lead to a subarachnoid hemorrhage, a type of stroke that is often fatal or severely disabling. Treatment can be either surgical, through clipping, where a metal clip is placed at the base of the aneurysm to block blood flow, or endovascular, using coiling techniques to fill the aneurysm and promote clotting from within.

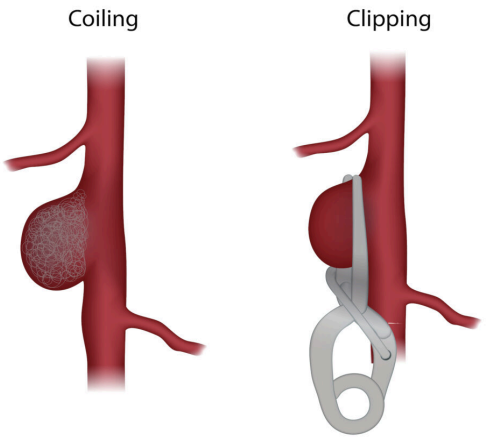


Figure 6 - Scheme of the coiling and clipping technique for aneurysm treatment

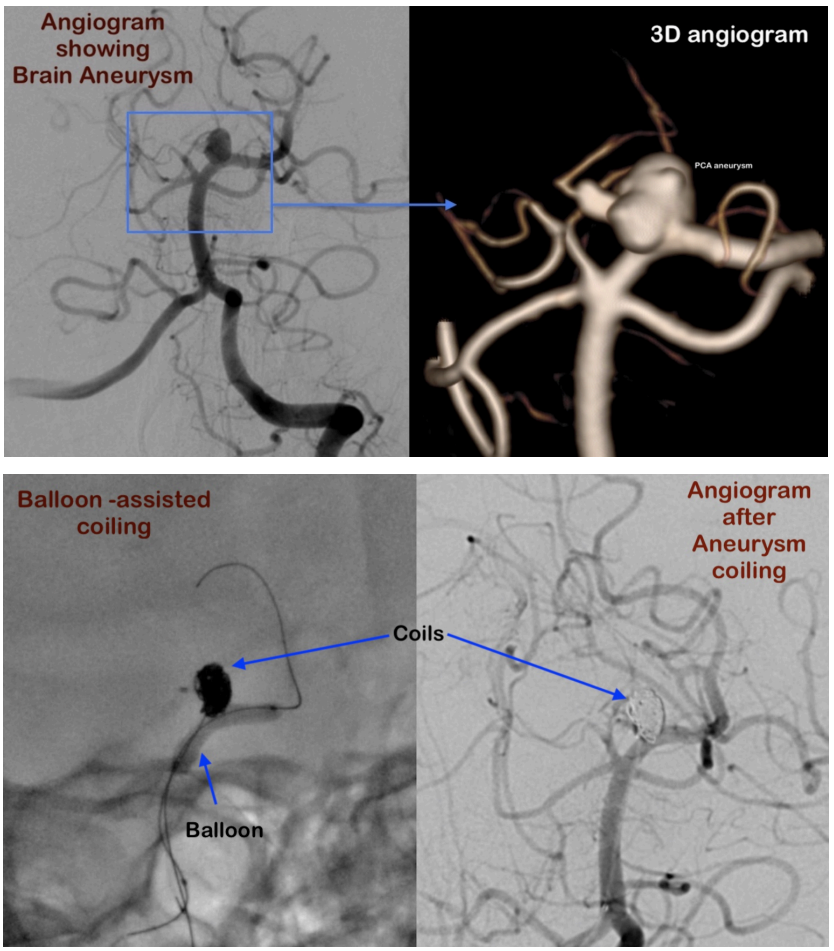


Figure 7 - Aneurysm coiling - case illustration

While these interventions are effective, they often require intensive post-operative monitoring, particularly due to the risk of complications such as rebleeding, vasospasm, hydrocephalus, or delayed ischemia. Manipulation of the brain during surgery can cause local inflammation in the area or disruption of cerebral circulation, thus provoking post-operative delirium, which can be exacerbated by certain factors such as age or significant psychological stress. Patients recovering from such procedures are frequently transferred to the intermediate care unit once their condition is stable enough to leave the intensive care, but still requires continuous observation. In this setting, patients may exhibit a range of neurocognitive symptoms, including confusion, reduced awareness, emotional instability, agitation, and cognitive fatigue, all of which can affect their behavior and increase their risk of sudden deterioration. These manifestations are often unpredictable and can vary greatly between individuals and over time. Pain is expressed in a variety of ways, ranging from subtle facial expressions to large, abrupt gestures, which makes real-time surveillance a critical and complex aspect of post-operative care.

1.5.3. Team organization

The project was led by a team of four students, each of whom had a defined role and a specific contribution. It is important to note that the technical aspects developed in this thesis correspond only to my own work.

As the only member of the group completing a Master's thesis, my responsibilities extended beyond technical contributions to include the overall coordination of the group, task distribution, and communication with the medical staff. Task allocation was carefully managed to leverage individual strengths: one student specialized in audio signal processing, another in ECG data handling, and a third in the implementation of Raspberry Pi-based data acquisition units.

The work developed in this thesis corresponds to the fine-tuning of the MoViNet-A5 model and the development of a multimodal attention mechanism. This management strategy not only facilitated smooth collaboration but also strengthened the team's ability to address practical challenges in a hospital setting.

2. State of the Art

2.1. Deep learning

2.1.1. Foundations and applications

Deep learning has emerged as one of the most transformative paradigms in modern artificial intelligence, offering state-of-the-art performance across a wide range of tasks, including computer vision, speech recognition, and natural language processing [3]. At its core, deep learning refers to a family of machine learning techniques based on artificial neural networks (ANNs) with multiple hidden layers, designed to automatically learn hierarchical representations of data. Unlike traditional machine learning models that rely heavily on handcrafted features, deep neural networks extract increasingly abstract and complex features directly from raw inputs, such as images, video, or audio signals [4].

Mathematically, a neural network is composed of layers of interconnected nodes (neurons), where each connection is associated with a weight w. Given an input vector x, a neuron computes a weighted sum $z = \sum_i w_i x_i + b$, where b is a bias term. This sum is then passed through a non-linear activation function, producing the neuron's output. By stacking many such layers, a deep neural network is able to model highly non-linear relationships between input and output data. The network's parameters are optimized through backpropagation, which computes the gradient of the loss function with respect to each weight and updates them iteratively using stochastic gradient descent or its variants [5].

One of the key strengths of deep learning lies in its ability to scale with large amounts of data and computational power. Landmark achievements such as AlexNet for image classification [6], ResNet for very deep architectures [7], and Transformer-based models for sequence modeling [8] have demonstrated that with sufficient training data, deep networks can outperform traditional methods by large margins. In video understanding specifically, 3D convolutional neural networks (3D CNNs) and transformer architectures have become dominant approaches, capable of capturing both spatial and temporal patterns [9].

In healthcare, deep learning has already shown promise in domains such as medical imaging diagnostics, ECG classification, seizure detection, and patient monitoring [10], [11]. Its ability to learn complex spatiotemporal dependencies is particularly relevant for continuous monitoring scenarios, where patient conditions may evolve dynamically over time and subtle visual or physiological cues can indicate critical deterioration. However, the deployment of deep learning in clinical environments faces challenges, including limited

annotated datasets, ethical concerns about transparency and privacy, and the high computational demands of modern architectures [12].

Overall, deep learning provides the theoretical and methodological foundation for this thesis. By leveraging pre-trained models and fine-tuning them on domain-specific data, it becomes possible to adapt powerful architectures to the healthcare context while respecting the constraints of real-time monitoring and clinical applicability.

2.1.2. Deep learning for action recognition

Action recognition in video has been one of the most challenging and dynamic fields in computer vision over the past decade. Unlike static image classification, which focuses solely on spatial features, action recognition requires understanding how objects and people evolve over time. This temporal dimension introduces complexity but is also critical in clinical monitoring, where subtle sequences of movements, such as a patient shifting repeatedly in bed before attempting to get up, may carry more significance than any single frame [13], [14].

The first generation of deep learning models for video relied on 2D convolutional neural networks (CNNs) applied frame by frame [15]. These models, originally developed for image recognition tasks, process spatial features effectively but ignore temporal dynamics, since each frame is analyzed independently. In practice, this often led to models that could recognize static postures but failed to capture transitions or actions, such as distinguishing between a patient sitting calmly and a patient preparing to leave the bed.

To address this limitation, researchers introduced 3D convolutional neural networks (3D CNNs), which extend the convolutional operation into the temporal dimension. While a 2D convolution computes feature maps by sliding a kernel W of size $(k_h \times k_w)$ over an image $X \in \mathbb{R}^{H \times W}$, a 3D convolution applies a kernel $W \in \mathbb{R}^{k_t \times k_h \times k_w}$ to a video clip $X \in \mathbb{R}^{T \times H \times W}$, where T is the number of frames. The resulting feature map is given by:

$$Y_{t,i,j} = \sum_{\tau=1}^{k_t} \sum_{m=1}^{k_h} \sum_{n=1}^{k_w} W_{\tau,m,n} \cdot X_{t+\tau,i+m,j+n}$$

This formulation enables the network to jointly learn spatial and temporal dependencies, making it far better suited for video understanding [16], [17]. Early models such as C3D [49] demonstrated the power of 3D convolutions, achieving significant improvements over 2D approaches on benchmark datasets.

Another important development came with two-stream networks [14], which introduced the idea of separating spatial and temporal information. One stream processes raw RGB frames to capture appearance, while the other processes optical flow fields to capture motion. The outputs are then fused to improve action recognition accuracy. This approach highlighted the importance of explicitly modeling temporal cues beyond static frame analysis.

More recently, models such as SlowFast networks [18] refined these ideas by using two parallel pathways: a "slow" pathway operating on sparsely sampled frames to capture semantic information, and a "fast" pathway operating at a higher frame rate to capture motion dynamics. This multi-rate design proved highly effective, particularly for fine-grained action recognition tasks, but it also increased computational demands, making deployment on resource-constrained devices more challenging.

Despite these advances, one of the recurring issues with 3D CNNs is their computational cost. Compared to 2D CNNs, the number of parameters and floating-point operations (FLOPs) increases significantly, resulting in high memory consumption and slow inference speed. This trade-off between accuracy and efficiency is especially relevant in clinical applications, where models must not only perform well but also run in real time on modest hardware. As will be discussed in Section 2.5, this challenge motivated the development of more efficient architectures such as MoViNet [19], which integrates innovations like stream buffering and causal convolutions to reduce memory bottlenecks without sacrificing accuracy.

In summary, deep learning for action recognition has progressed from simple frame-based 2D CNNs to sophisticated 3D CNNs and multi-stream architectures capable of modeling both spatial and temporal dynamics. These developments provide the foundation for applying AI in patient monitoring, where understanding actions such as agitation, attempts to leave the bed, or removal of medical devices requires robust temporal modeling. The challenge remains to adapt these powerful but computationally heavy models to the strict efficiency, interpretability, and reliability requirements of hospital environments.

2.1.3. CNNs, RNNs, and transformers

Deep learning for video understanding has historically been structured around three main paradigms: convolutional neural networks (CNNs), recurrent neural networks (RNNs), and more recently, transformers. Each paradigm reflects a different way of modeling spatiotemporal dependencies, with their own strengths and limitations [14], [16], [18].

CNNs remain the cornerstone of visual representation learning. Originally designed for image classification, CNNs apply convolutional kernels to extract local spatial features such as edges, textures, or object parts. In the video domain, these networks can either process frames independently (2D CNNs) or jointly across space and time (3D CNNs). As described earlier, 3D convolutions extend kernels into the temporal dimension, enabling direct modeling of motion.

While CNNs are highly effective at learning local spatiotemporal correlations, they struggle with long-term temporal dependencies, since the receptive field grows slowly with depth. This makes it challenging to capture actions that unfold over many seconds or minutes, as is often the case in clinical monitoring.

To address this limitation, RNNs and their variants, particularly Long Short-Term Memory (LSTM) networks [20], were introduced for sequential modeling. Unlike CNNs, RNNs process data step by step, maintaining a hidden state that summarizes past information.
LSTMs improve on RNNs by introducing gating mechanisms that regulate the flow of information and alleviate the vanishing gradient problem. This enables them to capture longer-term temporal dependencies, making them attractive for early action recognition research. However, RNNs are inherently sequential in their computation, which limits parallelization and makes training on large-scale video datasets inefficient.

The emergence of transformers [21] revolutionized sequence modeling by replacing recurrence and convolution with self-attention mechanisms. In this framework, each input element attends to every other element in the sequence, weighted by learned attention scores.

For vision tasks, this concept was adapted in the Vision Transformer (ViT) [22], which splits an image into patches treated as tokens, analogous to words in natural language. For videos, spatiotemporal transformers [23], [24] extend this idea by embedding both spatial and temporal patches, enabling the model to learn complex motion patterns across long sequences. The use of multi-head attention further enhances capacity by allowing the model to jointly capture different types of dependencies.

Transformers offer several advantages over CNNs and RNNs: they are fully parallelizable, scale effectively with data, and can capture both short- and long-range dependencies in a single framework. However, these benefits come at the cost of computational and memory intensity, especially when applied to long video sequences. This makes them powerful benchmarks for action recognition accuracy but less suited for real-time deployment in resource-constrained settings such as hospital monitoring systems.

2.2. Video-based patient monitoring

In recent years, the issue of video surveillance of patients in the specific context of hospital environments has become increasingly important, as monitoring is often essential for the patient's safety but also synonymous with a heavy workload for medical staff. Traditionally, patient supervision relied on a combination of bedside observation and physiological sensors such as ECG, pulse oximeters, or blood pressure monitors. While effective for tracking vital signs, these systems do not capture visible patient behaviors, such as agitation, attempts to leave the bed, or removal of catheters, which are often the earliest indicators of distress or risk [26], [27]. For this reason, video-based continuous monitoring systems have emerged as a complementary solution, aiming to ensure the patient's safety by observing his activity in real time.

Numerous research groups have explored the use of computer vision for the purpose of analysing hospital videos. Early approaches relied on conventional image processing methods such as background subtraction, motion tracking, and handcrafted feature extraction [28]. These techniques made it possible to recognise typical movements such as patients getting out of bed or falling, but lacked robustness in the context of complex clinical environments, where lighting conditions, occlusions, and subtle change in facial expressions make detection more challenging. With the rise of deep learning, especially convolutional neural networks (CNNs), more sophisticated methods have been proposed to automatically extract features from raw video and classify patient states with higher accuracy [29].

The flagship application of patient monitoring in hospitals is fall detection. Falls remain one of the most common and easily detectable events in healthcare, particularly among elderly or post-surgical patients. The aim of patient monitoring in this context is to prevent these falls and thus reduce their occurrence. Studies have shown that video systems combined with deep learning can detect attempts to exit the bed, restlessness, or sudden collapses before an actual fall occurs [30]. In some implementations, alarms are triggered in real time,

enabling staff to intervene quickly and prevent injury [31]. Beyond fall detection, detecting episodes of delirium and agitation, signs of suffering that are especially common in intermediate care units after neurological procedures, is another application for which video surveillance is proving relevant. Such states are characterized by restless movements, attempts to remove catheters, or facial expressions of pain, which can be captured more reliably by cameras than by physiological monitors alone [32].

Despite these advances, several limitations remain. A first major issue is the quality and availability of datasets. To train effective models, large-scale annotated datasets are required. However, due to privacy concerns, very few hospitals have been able to collect and release real patient recordings. Many published systems therefore rely on simulated datasets, using actors or medical staff to reproduce typical patient movements [26], [29]. While useful for proof-of-concept demonstrations, such datasets often fail to capture the diversity and unpredictability of real patients, leading to reduced generalizability in practice. In the context of the present thesis, this limitation was also encountered, as simulated recordings had to be used for initial validation.

Another challenge relates to anonymization. In traditional video analysis, anonymization can be achieved by blurring faces or masking backgrounds. However, in the medical context, facial expressions often carry essential clinical information. Subtle signs of pain such as grimacing, eye movements, or head shaking may indicate confusion, or distress [19]. If these features are removed during anonymization, much of the clinically relevant signal is lost, thereby limiting the system's effectiveness. This raises ethical concerns that must be carefully managed [34].

Finally, video-based monitoring must also account for workflow integration. A system that generates too many false alarms may quickly become unusable due to alarm fatigue, a well-documented problem in intensive and intermediate care units [9]. At the same time, a system that misses critical events (false negatives) would fail in its most important function: ensuring patient safety. As such, the key challenge is to find a balance between sensitivity and specificity, while designing tools that help and support nurses and assistant nurses in their work, rather than making them even more tired.

In summary, video-based patient monitoring represents a promising extension of current clinical supervision methods. It enables the detection of complex, non-physiological signs of distress, provides real-time support for specific movement detection , and reduces the need for continuous human observation. However, its effectiveness depends on the quality of training datasets, the handling of privacy concerns, and its integration into clinical workflows. These challenges highlight the importance of developing robust, ethically sound, and clinically validated systems, forming the foundation for the present thesis.

2.3. Multimodal monitoring

Although video surveillance is a powerful tool for observing patient behaviour, it is clear that relying on a single modality has certain limitations. Many clinical situations cannot be fully detected by visual recognition alone, as they do not necessarily manifest themselves through visible movements. Certain respiratory patterns or cardiac activities may change, and sounds may be emitted. Thus, certain pains or signs that are cause for concern regarding the patient's health may be expressed through more subtle physiological signs. This has led to a growing body of research on multimodal monitoring systems, which combine video with other data sources to provide a more comprehensive and reliable view of patient status [36], [32].

Among the most common complementary modalities are audio signals and physiological measurements. Audio can capture patient vocalizations such as moaning, shouting, or calls for help, which often precede physical agitation or deterioration. Research has shown that audio-based features, when combined with video, can significantly improve classification accuracy in patient monitoring tasks [25]. Likewise, physiological signals, particularly from the Electrocardiogram (ECG), offer direct insight into cardiac and respiratory states. In hospital monitoring systems, ECG and related measurements such as blood pressure or oxygen saturation are already continuously available through the Patient Monitoring System (PMS). By integrating these signals with video-based action recognition, it becomes possible to detect not only what the patient is doing, but also how the patient is physiologically responding [32], [37].

The main challenge of multimodal monitoring lies in the weight given to each channel. For example, if the patient remains calm in bed, showing stable heart rate and blood oxygenation levels, the system will predict that no intervention by nurses is necessary. Conversely, if even the slightest movement is combined with an alarming sound and abnormal physiological data, the system will immediately detect a need for intervention. This motivates the use of multimodal attention mechanisms [25], [29], which allow the model to dynamically adjust the importance of each modality depending on the context. In practice, this means that video, audio, and physiological data are processed separately through specialized feature extractors before being fused at a later stage. The attention mechanism then learns, during training, to assign greater weight to the most informative modality for each situation.
Another challenge of multimodal attention is its time cost. Indeed, a system of this type is difficult to operate in real time due to its complexity. Although it offers a robust alternative to the detection of specific events, its implementation in real time remains difficult to imagine.

In summary, multimodal surveillance is a promising component of automated surveillance. By combining different types of input such as sound, heart rate, and video, these models are well positioned to detect complex patient's states, to handle uncertain or noisy conditions, and to provide more reliable alerts. This line of research directly addresses one of the main challenges of hospital AI: moving from isolated signal processing toward context-aware monitoring systems capable of supporting clinical decision-making in real time.

2.4. Ethical and privacy considerations in healthcare AI

The use of artificial intelligence for continuous patient monitoring raises a number of ethical and legal questions that must be addressed before any clinical implementation. At the European level, the regulation of video processing and physiological data is done by the General Data Protection Regulation, which mandates principles such as lawfulness, purpose limitation, data minimization, storage limitation and strict access controls. In short, data controllers must have a lawful basis for processing, such as explicit patient consent, implement appropriate technical and organisational safeguards, and ensure that personal data are retained only as long as necessary. [47]

Beyond the legal texts, a recurring practical tension in video-based clinical research is the trade-off between privacy (anonymization) and clinical utility. Many standard approaches to privacy, like blurring faces or masking identifying regions for instance, reduce the risk of re-identification but also remove clinically relevant signals such as facial micro-expressions, grimacing or subtle head movements that are strong indicators of pain, agitation or delirium [39].

As a consequence, some ethically-approved medical studies have permitted access to non-anonymized video under strict governance (limited, logged access and short retention), because anonymization would materially undermine the scientific and clinical value of the recordings [26].
Informed consent, transparency and patient autonomy are therefore central. Patients must be given clear, accessible information about what is recorded, for what purpose, who will access the recordings, how long the data will be kept, and how they can withdraw consent. In hospital implementations, transparent communication with both patients and staff is also essential to build trust and acceptance.

When ethical approval allows viewing of non-anonymized video for model training, governance must be strict and technically enforced. Best practice includes: minimizing the number of individuals with raw-video access, storing videos on encrypted, hospital-managed storage, keeping identifiers separate from clinical recordings, deleting personally identifying metadata as soon as possible, and deleting or archiving data according to a predefined retention schedule. Many hospital projects adopt a limited-access model and explicit, auditable policies for local storage and backup; such practices were recommended in pilot implementations and reviews of continuous video monitoring systems [40], [35].

The measures put in place for data backup must be commensurate with the sensitivity of the data. Among these, the GDPR and associated national bodies refer to encrypted backups, strict user authentication, regular audits, and defined breach-response procedures as necessary prerequisites for any hospital AI deployment [47].
Algorithmic fairness, bias and model transparency must also be considered as ethical issues in their own right. Models trained on small or non-representative datasets risk performing poorly for under-represented patient groups, which can worsen inequities in care. The literature highlights the need for continuous evaluation across demographic and clinical subgroups, re-training with new real-world data. Clinical acceptance studies further show that staff trust increases when models provide interpretable outputs and when clinicians are involved in iterative validation workshops [41], [42].

From a clinical-workflow perspective, ethical deployment also means designing the system to support rather than replace clinical judgment. Systems should be validated in parallel with standard care before any automated alerting is trusted; alarm thresholds and the balance between sensitivity and false alarms must be tuned together with nursing staff to avoid alarm fatigue [35], [28].

2.5. Summary of research gaps

Despite the significant progress in applying AI to clinical video monitoring, several research gaps remain that limit both scientific understanding and practical deployment in real hospital environments. First, most existing studies have been performed on simulated datasets or controlled experimental environments rather than on real patient populations [32], [37], [40]. This limitation affects the generalizability of results, as real hospital settings present higher variability in lighting, patient positioning, and background activity. Moreover, few studies systematically address the difficulty of building ethically-approved,

large-scale clinical datasets, which is still one of the main bottlenecks for robust training and validation [26], [43].

Another major gap concerns the multimodal integration of heterogeneous data. While video-based recognition has been widely studied, the combined use of video, audio, and physiological signals remains underexplored [35]. Most architectures either focus exclusively on visual data or handle additional modalities in a simplistic, late-fusion manner. There is limited evidence on how advanced mechanisms such as multimodal attention or spatiotemporal transformers perform when combining complex clinical data sources in practice. This lack of exploration represents a missed opportunity, since multimodal fusion could provide greater robustness and reduce false negatives in high-risk monitoring tasks.

Another open question is robustness to clinical variability. Existing studies often report high accuracy but rarely evaluate how models generalize to different hospitals, patient groups, or recording conditions [36], [44]. There is a pressing need for multi-center datasets and external validation to ensure that AI systems do not overfit to one hospital's specific practices or environment. Similarly, bias and fairness issues remain understudied, as most datasets are too small to capture demographic diversity [28].

Finally, ethical and clinical adoption challenges remain insufficiently researched. While several works emphasize privacy-preserving techniques or anonymization methods, there is little empirical evidence on how these approaches affect both model performance and staff acceptance [27]. Furthermore, few studies investigate how to integrate AI monitoring into the existing workflow without increasing alarm fatigue or cognitive burden on nurses [35]. Research that combines technical evaluation with human-centered design and clinical workflow studies is therefore essential to bridge the gap between promising laboratory results and actual patient benefit.

3. Materials and methods

3.1. Ethical approval and patient consent

This study is conducted within the framework of an ethical approval granted by the Swedish Ethical Review Authority, which regulates all aspects of data collection, processing, and storage. The approval covers the acquisition of multimodal patient data, including video recordings (with sound) from room-installed cameras, physiological signals from the Patient Monitoring System (ECG, blood pressure, pulse, oxygen saturation, temperature, respiratory rate), and field observations. To label relevant moments for AI training, each video stream is marked in real time using a physical button connected to the camera system. In parallel, nurses or assistant nurses record timestamps in a written logbook, creating a dual-tagging mechanism.

Crucially, the collected video data is not anonymized, as the objective is to train machine learning models on realistic behavior in clinical environments. However, all patient identifiers are kept separate from the video material. A temporary list containing social security numbers may be stored exclusively by a clinical staff member, solely for the purpose of conducting patient follow-ups after discharge. This list is never shared with researchers or developers and is deleted immediately after use. It is not part of the research data itself.

Workshops involving nurses, assistant nurses, and technology developers are also part of the approved research activities. These include the analysis of anonymized video clips for model calibration, as well as collaborative sessions for system design. Access to the collected data is restricted to a core team of six authorized personnel: two researchers from KTH, one from Karolinska Institutet, one from Karolinska University Hospital, and two technology developers. No other parties are granted access. The data will be retained (in anonymized form) until 2030 for research purposes and cannot be repurposed outside the scope of this project. This framework ensures that all procedures align with ethical and legal requirements, while enabling meaningful development of clinical AI systems [Appendix 3].

One of the key issues in the project concerned patient consent. Patients involved in the project are prone to periods of delirium during which they are not considered to be psychologically capable of giving consent. It is therefore necessary to inform patients about the project and ask for their consent before they are admitted to intermediate care. However, given the timing of these patients' hospital stay, it is delicate to ask for their consent to be filmed during difficult moments of their life, while they are being told that they need to

undergo an intracranial surgery. These announcements can be quite overwhelming, therefore, empathy should always remain at the heart of the project.
This issue was the subject of much debate among the medical teams involved, and it was finally decided that doctors would discuss the project with patients and ask for their consent during explanatory meetings about the surgery they were going to undergo. This means that patients' consent is obtained several weeks before they enter the intermediate post-surgery care unit.

3.2. Data collection setup

3.2.1. Camera system and Raspberry Pi setup

The camera setup was implemented using three Raspberry Pi boards connected to 3D-printed mounts. Each Pi handled one video stream and synchronized its internal clock with the tagging system. This lightweight and modular infrastructure was designed for scalability across different patient rooms. The choice of Raspberry Pi was also motivated by its ability to support lightweight real-time AI inference models, such as the one developed in this study.

Figure 8 - Cameras and Raspberry Pi setup

The electronic infrastructure of the monitoring system was designed to be both reliable and unobtrusive within the clinical environment. The setup consists of multiple cameras connected to Raspberry Pi devices through Ethernet switches, supported by a simple control unit containing an on/off switch, a tagging push button, and indicator LEDs. Video data is stored locally on Raspberry Pis and backed up on a 1 TB external hard drive. To minimize the visibility of the system, the Raspberry Pis are placed in the false ceiling while cameras, equipped with 12 MP sensors at 30 FPS, are mounted on custom 3D-printed supports adapted to the rails of hospital ceilings. The entire system is powered safely through the hospital's electrical infrastructure, using 5V/3A USB-C inputs, ensuring protection against misconfigured power delivery.

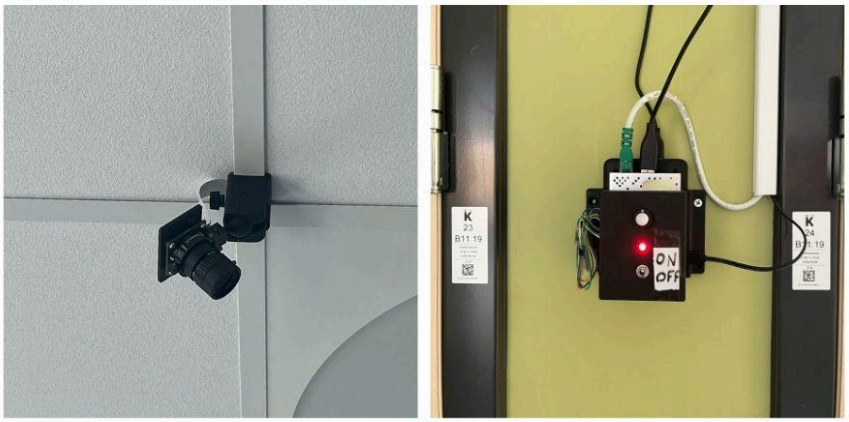


Figure 9 - Camera and button box in the hospital environment

Special attention was given to usability, as nurses and assistant nurses work under stressful conditions and must interact with the system seamlessly. For this reason, the tagging button was installed at the entrance of each patient room, allowing staff to press it quickly before entering for an intervention. The user interface was intentionally designed to remain minimal and intuitive. The future alerting interface will also be co-developed with medical staff to ensure that its features align with their workflow.

In terms of performance, the hardware is required to operate continuously and withstand environmental variations such as changes in light, and noise. However, one limitation identified was luminosity in hospital rooms at night: the cameras used in this setup cannot capture usable video in low-light conditions. For this reason, the scope of this thesis was restricted to analyzing daytime movements only, while the development of solutions for nighttime monitoring is left for future research (see Section 5.4. Suggestions for Future Work).

Placement of devices was carefully planned to avoid obstructing medical operations or posing risks of falling. This lightweight and modular design makes the infrastructure scalable across different rooms while maintaining robustness and safety standards expected in a hospital environment.

3.2.2. Control box and labeling mechanism

In the initial phase, we focused on collecting video data only. Three cameras were installed in a patient room, capturing footage from different angles. Using multiple viewpoints is particularly important for training a robust model, as patient movements may be partially occluded or difficult to interpret from a single perspective. In our case, this setup was also essential to capture patients' facial expressions, such as grimaces, which can signal pain or distress and represent subtle but clinically meaningful cues for intervention. One of the cameras was therefore positioned above the patient's bed to ensure these nuances could be recorded. For this reason, building an anonymized dataset with blurred patient faces would have been impractical in this project, as it would have eliminated crucial information related to patients' facial expressions and overall behavior.

A control box was placed at the room's entrance, featuring two main components: a power switch, allowing staff to stop video recording during privacy-sensitive moments, and a push-button to tag video frames in real time when a nurse or assistant entered the room to perform an intervention. The Raspberry Pi devices were programmed to associate each video stream with a binary labeling system: continuous frames were assigned a value of '0' when no intervention was needed, and a value of '1' was generated at the exact frame where the button was pressed, indicating that an intervention had begun. This simple binary scheme forms the basis of the supervised learning task: the model is trained to discriminate between periods requiring no intervention and those where an intervention is necessary.

However, assigning a "1" is not a trivial process. Unlike a purely mechanical event, the decision to intervene depends heavily on the clinical judgment and expertise of the nurses, which may vary depending on context, patient behavior, and professional experience. There is no universal or predefined threshold that clearly distinguishes when an intervention becomes necessary. This introduces an inherent subjectivity in the labeling process, underscoring the importance of close collaboration with nurses and assistant nurses. Their input is essential not only for ensuring accurate labeling but also for defining clinically meaningful criteria that the AI system should ultimately learn to recognize.

Figure 10 - Patient room equipped with the camera setup in Karolinska, Solna

To ensure accuracy during the simulation phase, all collected videos were manually reviewed in fast-forward mode, and missing or misplaced labels were corrected. In preparation for future real-world data collection, a complementary mechanism was designed: nurses will be able to report missed button presses using a dedicated log sheet, where approximate times of interventions can be noted [Appendix 2]. While this procedure was not yet implemented during the current simulations, it will support the creation of a more reliable and clinically validated dataset in future stages of the project.

3.3. Dataset construction and preprocessing

3.3.1. Simulation data collection

Due to the late arrival of patient consent and logistical constraints, the full dataset of real patient recordings could not be collected in time for training. As an alternative, we conducted simulation sessions inside the hospital room using the complete monitoring setup to generate labeled video data. Over a continuous period of approximately five hours, three different individuals took turns simulating patient behavior in the room, reproducing a variety of movements and actions that could occur in real clinical contexts. This approach ensured diversity in body types, gestures, and reactions, providing a broader basis for testing the system.

Figure 11 - Examples of frames associated to a tag ‘1’ (intervention required)*

Figure 12 - Examples of frames associated to a tag ‘0’ (no intervention required)*

* the pictures have been voluntarily blurred and discoloured in this report for privacy reasons

The simulation served multiple purposes. First, it allowed us to validate the hardware infrastructure, including the three-camera setup, the control box with its push-button tagging mechanism, and the Raspberry Pi data storage and synchronization. Second, it enabled us to test the end-to-end pipeline, from video capture and labeling to dataset construction and preliminary training of the models. Importantly, the simulation confirmed that the system was capable of running continuously without major failures, which is a key requirement for its eventual clinical deployment.

Although the simulated dataset is not fully representative of real patient behavior, since it lacks the unpredictability, emotional cues, and clinical context of actual hospital cases, it nevertheless provided a functional and practical testbed for both the technical infrastructure and the AI models. Some limitations and potential improvements were identified during this phase (see Section 4.3. Discussion), but overall the sessions proved successful in demonstrating feasibility. The simulated data also prepared the ground for future student teams, who will be able to continue development using actual patient datasets once consent and logistical conditions are fully met.

3.3.2. Preprocessing pipeline and data augmentation

Before using the dataset for training, a thorough post-processing and verification step was required. One of the main challenges encountered was the latency between the push-button tags and the actual video frames, which caused discrepancies between the intended intervention moments and their recorded timestamps, the possible cause being a cable with poor throughput installed in the hospital room.
To address this, all videos were reviewed in fast-forward mode, and each tag was manually verified and corrected when necessary. This manual curation ensured that the dataset was as reliable and representative as possible for training the AI system.

The dataset creation process was built upon a collection of raw video recordings, each associated with an annotation file. To facilitate synchronization and dataset construction, a dedicated CSV file was created to systematically organize the data. This file recorded for each sequence the initial frame, relative frame indices, and the corresponding binary tags (0 = no intervention, 1 = intervention). This structure allowed for easier parsing, clip segmentation, and dataset management, while providing a transparent mapping between video data and intervention labels.

The use of relative frames was particularly beneficial. It not only helped to counterbalance the latency issues observed during tagging but also significantly reduced runtime operations during training, since clips could be pre-indexed and accessed efficiently without scanning entire video files. This improvement made the training process faster and more stable, even on limited hardware.

	Raw videos	Clips generated	Batches	
			Train dataset	Test dataset
Quantity	5	1440	360	90
Length (of each)	1 hour	2-3s	4 to 16 frames per batch (depending on the model)	

Table 1 - Simulated dataset distribution

The datasets were constructed using TensorFlow's tf.data API, which leveraged the .csv metadata to stream clips efficiently. However, initial versions of the pipeline encountered multiple technical challenges. Real-time frame extraction caused excessive RAM usage and prolonged loading times, often resulting in training session interruptions.

Finally, to increase dataset variability and enhance model generalization, several data augmentation techniques were applied to the video clips. These included horizontal flipping, random cropping, and brightness adjustments, which simulate the variability of real hospital conditions such as different lighting environments, patient positions, or camera perspectives. By introducing such diversity, the augmented dataset helps mitigate overfitting and strengthens the robustness of the trained models.

3.3.3. Guided sampling strategy

In order to further address the issue of class imbalance and improve the representativity of rare action categories during training, a guided sampling strategy was employed. Rather than uniformly sampling video clips, the data loader was modified to prioritize the selection of underrepresented classes. This was achieved by computing class frequencies in the training set and assigning higher sampling probabilities to minority classes.

By integrating guided sampling, the model was exposed to a more balanced distribution of examples during each epoch, enhancing its ability to learn decision boundaries for less

frequent actions. This method proved particularly useful in the context of limited dataset diversity, where traditional random sampling would otherwise reinforce existing imbalances.

3.4. Model Architectures

This project investigates and compares two main deep learning architectures: MoViNet and PredFormer.

3.4.1. MoViNet

MoViNet (Mobile Video Network) is a family of deep learning models specifically designed for efficient and real-time video action recognition. Unlike traditional architectures that often require powerful GPUs and batch processing of long video clips, MoViNet was developed to support continuous streaming inference on lightweight hardware, making it suitable for real-time applications such as patient monitoring [28].

At its core, MoViNet builds upon 3D convolutional neural networks (3D-CNNs), which extend standard 2D convolutions to the temporal dimension as previously discussed in the state of the art section. However, full 3D convolutions are computationally expensive and memory-intensive.

To mitigate this, MoViNet adopts factorized convolutions (2+1D), decomposing a 3D kernel into a spatial convolution followed by a temporal one:

$$Y(t,i,j) = \left(\sum_{m,n} X(t, i+m, j+n) \cdot K_s(m,n) \right) * K_t(p),$$

where K_s is the 2D spatial kernel and K_t the 1D temporal kernel. This reduces the number of parameters while retaining the ability to model spatiotemporal features.

A central innovation of MoViNet is the temporal stream buffer, which enables continuous inference without recomputing all activations for each new clip. Instead of processing overlapping subclips independently, intermediate feature maps at subclip boundaries are

cached and reused. If h_t denotes the hidden activation at time t, the buffer stores h_t so that the next prediction uses:

$$h_{t+1} = f(X_{t+1}, h_t),$$

where f is the convolutional transformation. This reduces memory usage from $O(T)$ to $O(1)$, with reported savings of up to 90% for MoViNet-A5. Importantly, only causal operations are used, meaning that predictions depend on past and current frames, but never on future ones. This property makes MoViNet viable for real-time monitoring in clinical environments.

Beyond its streaming capability, MoViNet integrates architectural elements from MobileNetV3, such as inverted residual blocks and squeeze-and-excitation (SE) modules. The latter act as channel-wise attention mechanisms. It dynamically reweights feature maps to emphasize informative channels while suppressing irrelevant ones.

MoViNet models are available in different scales (A0–A5), balancing efficiency and accuracy. Smaller models (A0, A1) are highly efficient for mobile devices, while larger ones (A4, A5) achieve state-of-the-art accuracy. In this thesis, MoViNet-A5 was selected as it provides the best compromise for our application: accurate recognition of patient behaviors while still being deployable on relatively constrained hardware.

By combining 2+1D convolutions, causal buffering, and lightweight attention modules, MoViNet achieves both efficiency and robustness, making it particularly well-suited for real-time monitoring tasks in hospital environments.

STAGE	OPERATION	OUTPUT SIZE
data	stride 5, RGB	120×320^2
conv_1	1×3^2, 24	120×160^2
block_2	$\begin{bmatrix} 1\times5^2, 24, 64 \\ 1\times5^2, 24, 64 \\ 3\times3^2, 24, 96 \\ 3\times3^2, 24, 64 \\ 3\times3^2, 24, 96 \\ 3\times3^2, 24, 64 \end{bmatrix}$	120×80^2
block_3	$\begin{bmatrix} 5\times3^2, 64, 192 \\ 3\times3^2, 64, 152 \\ 3\times3^2, 64, 152 \\ 3\times3^2, 64, 152 \\ 3\times3^2, 64, 192 \\ 3\times3^2, 64, 192 \\ 3\times3^2, 64, 192 \\ 3\times3^2, 64, 152 \\ 3\times3^2, 64, 152 \\ 3\times3^2, 64, 192 \\ 3\times3^2, 64, 192 \end{bmatrix}$	120×40^2
block_4	$\begin{bmatrix} 5\times3^2, 112, 376 \\ 3\times3^2, 112, 224 \\ 3\times3^2, 112, 376 \\ 3\times3^2, 112, 376 \\ 3\times3^2, 112, 296 \\ 3\times3^2, 112, 376 \\ 3\times3^2, 112, 224 \\ 3\times3^2, 112, 376 \\ 3\times3^2, 112, 376 \\ 3\times3^2, 112, 296 \\ 3\times3^2, 112, 376 \\ 3\times3^2, 112, 376 \\ 3\times3^2, 112, 376 \end{bmatrix}$	120×20^2
block_5	$\begin{bmatrix} 5\times3^2, 120, 376 \\ 3\times3^2, 120, 376 \\ 3\times3^2, 120, 376 \\ 3\times3^2, 120, 376 \\ 1\times5^2, 120, 224 \\ 3\times3^2, 120, 376 \\ 3\times3^2, 120, 376 \\ 3\times3^2, 120, 224 \\ 3\times3^2, 120, 376 \\ 3\times3^2, 120, 376 \\ 3\times3^2, 120, 376 \end{bmatrix}$	120×20^2
block_6	$\begin{bmatrix} 5\times3^2, 224, 744 \\ 3\times3^2, 224, 744 \\ 1\times5^2, 224, 600 \\ 1\times5^2, 224, 600 \\ 1\times5^2, 224, 744 \\ 1\times5^2, 224, 744 \\ 3\times3^2, 224, 744 \\ 1\times5^2, 224, 896 \\ 1\times5^2, 224, 600 \\ 1\times5^2, 224, 600 \\ 1\times5^2, 224, 896 \\ 1\times5^2, 224, 744 \\ 3\times3^2, 224, 744 \\ 1\times5^2, 224, 896 \\ 1\times5^2, 224, 600 \\ 1\times5^2, 224, 600 \\ 1\times5^2, 224, 744 \\ 3\times3^2, 224, 744 \end{bmatrix}$	120×10^2
conv_7	1×1^2, 992	120×10^2
pool_8	120×10^2	1×1^2
dense_9	1×1^2, 2048	1×1^2
dense_{10}	1×1^2, 600	1×1^2

Figure 13 - MoViNet-A5 architecture

3.4.2. PredFormer

PredFormer is a transformer-based architecture developed for video prediction tasks, and it distinguishes itself by removing both convolutional and recurrent operations. Instead, it relies solely on spatiotemporal self-attention to capture dependencies across space and time. This design gives PredFormer a global receptive field, which is crucial when modeling long-range temporal structures or subtle behavioral signs that cannot be efficiently captured by CNNs or RNNs. In the clinical monitoring context, such as detecting early agitation in patients, the ability to model gradual, temporally extended changes makes PredFormer an interesting candidate despite its high computational demands and lack of real-time feasibility [14].

Formally, given an input sequence of video frames

$$X = \{x_{t-T+1}, \ldots, x_t\},\ x_i \in \mathbb{R}^{H \times W \times C},$$

the goal of PredFormer is to learn a mapping

$$F_\Theta\colon X \to \hat{Y} = \{\hat{y}_{t+1}, \ldots, \hat{y}_{t+T}\},$$

where $\hat{y}_j \in \mathbb{R}^{H \times W \times C}$ are the predicted future frames. The model parameters Θ are optimized by minimizing a reconstruction loss such as the mean squared error

$$L(\hat{Y}, Y) = \frac{1}{T' \cdot H \cdot W \cdot C} \sum_{j=1}^{T'} \left\| \hat{y}_{t+j} - y_{t+j} \right\|_2^2,$$

with optional perceptual or adversarial terms to encourage sharper predictions.

Each video frame is divided into non-overlapping patches of size $p \times p$, flattened and projected through a linear embedding layer into a latent space of dimension D. The result is a sequence of tokens

$$Z_0 = \left[z_0^1, z_0^2, \ldots, z_0^N \right],\ z_0^i \in \mathbb{R}^D,$$

where $N = \frac{HW}{p^2}$ is the number of tokens per frame. To retain temporal order, PredFormer applies sinusoidal positional encodings to each token.

The backbone of PredFormer is the Gated Transformer Block (GTB), which modifies the standard transformer encoder by integrating gating mechanisms for improved control of information flow. Within each block, the multi-head self-attention mechanism computes:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

where queries Q, keys K, and values V are linear projections of the input tokens.

An important architectural choice concerns the structuring of attention. PredFormer can employ full joint spatiotemporal attention, but due to quadratic complexity in sequence length, factorized attention is often preferred. Therefore, attention is alternately applied across the temporal and spatial dimensions. For instance, a temporal attention block models dependencies across frames for each spatial location, followed by a spatial attention block across patches within each frame. Interleaving these blocks provides flexibility in balancing accuracy and efficiency. Such design is particularly useful in patient monitoring scenarios, where some risk behaviors appear abruptly while others evolve slowly.

The decoder is kept deliberately simple. Since the encoder already maintains spatiotemporal context. This lightweight reconstruction head reduces overhead while ensuring high-quality predictions.

PredFormer thus embodies the power of transformer-based architectures in video modeling. Its strength lies in the absence of strong inductive biases tied to locality or sequential processing, enabling it to learn subtle, long-range dependencies. While this comes at the cost of high memory consumption and limited real-time feasibility, its accuracy in offline settings makes it an excellent benchmark against lightweight architectures like MoViNet. In the context of this thesis, PredFormer serves to assess whether transformer-based models provide a meaningful advantage in recognizing complex or temporally extended patient behaviors, paving the way for future multimodal transformer frameworks in clinical AI.

3.4.3. Preliminary comparison

To assess the potential of the PredFormer architecture for clinical video monitoring, we conducted a preliminary comparison with MoViNet focusing strictly on two baseline criteria: processing speed and memory requirements. MoViNet is designed for low-latency inference and edge deployment, and benefits from aggressive model compression and quantization. It is capable of running efficiently on devices with limited computational power such as mobile processors or embedded systems, including Raspberry Pis in our case. In contrast, PredFormer is a transformer-based model that forgoes both convolution and recurrence in favor of self-attention mechanisms applied across spatiotemporal sequences. While this allows for superior modeling of long-range dependencies, it also introduces a higher computational overhead, especially during inference. The self-attention operation scales quadratically with the number of input tokens, making it memory-intensive and less suited to real-time performance on resource-constrained hardware. In practice, preliminary testing shows that PredFormer requires more VRAM (typically 6–12 GB) and has significantly slower inference speed, particularly on long video sequences or high-resolution frames.

Therefore, although PredFormer shows architectural promise in terms of modeling capacity, our analysis confirms that MoViNet remains more appropriate for real-time hospital deployment where low-latency and lightweight memory usage are critical. PredFormer could, however, be explored further in offline settings or as a benchmark model for future accuracy-focused evaluations.

Feature	MoViNet	PredFormer
Core mechanism	3D convolutions with stream buffering	Spatiotemporal self-attention (no conv/RNN)
Complexity	Linear in sequence length	Quadratic in sequence length
Memory usage	Low (optimized for edge devices)	High (requires powerful GPUs)
Training data requirement	Moderate (pre-trained on Kinetics-700)	Very large datasets (e.g., Kinetics, Human3.6M)
Role in this thesis	Deployable solution	Benchmark for recognition accuracy

Table 2 - Preliminary comparison between MoViNet and PredFormer

3.4.4. Multimodal integration and attention mechanism

Beyond video, the system was designed to incorporate audio and ECG signals. Each modality was processed separately before being fused via a multimodal attention mechanism, which dynamically weighs the contribution of each modality to the final classification. This integration was conducted experimentally, as real multimodal data could not be collected during the thesis period due to limited access.

As discussed in the state of the art section, multimodal attention is a mechanism used in deep learning models to effectively combine and interpret information coming from multiple sources, or modalities, such as video, audio, and physiological signals. Rather than treating all input data equally, attention mechanisms allow the model to dynamically focus on the most relevant features from each modality depending on the context.

In this project, each modality was first processed independently by a feature extraction module tailored to its nature. The video stream was passed through a spatiotemporal backbone (MoViNet or PredFormer), and fine-tuned. Audio signals were converted into spectrograms and encoded via a convolutional feature extractor. ECG waveforms were processed using a temporal encoder designed to capture rhythmic patterns in the signal. Each of these steps generated latent feature vectors of dimension d, which were then projected into a common embedding space to allow joint processing.

The integration of these features was achieved through a multi-head attention mechanism, a central element of transformer architectures. Similarly to the PredFormer model, for an input sequence of embeddings $\mathbb{Z} \in \mathbb{R}^{n \times d}$, queries Q, keys K, and values V are computed through learned linear projections:

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V$$

where W_Q, W_K, $W_V \in \mathbb{R}^{d \times d_k}$. The attention operation evaluates the relevance between tokens by computing scaled dot-products:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d^k}}\right)V$$

The multi-head replicates this operation h times with different learned projections, yielding:

$$MultiHead(Q, K, V) = Concat\left(head_1, \ldots, head_h\right)W_0$$

where each $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$. This mechanism enables the model to jointly attend to information from different representation subspaces at multiple scales. In practice, some heads may specialize in capturing short-term temporal correlations (e.g., rapid spikes in ECG or sudden noises), while others focus on long-range dependencies (e.g., gradual patient agitation visible in video).

The choice of multi-head attention in this project was motivated by its ability to dynamically weigh the relative importance of modalities depending on context. For instance, a sudden acceleration in heart rate might be insignificant in isolation, but when combined with unusual movements captured in video and abnormal sounds, it becomes a strong indicator of patient distress. A single-head mechanism would be limited in its representational capacity, whereas multi-head attention allows for richer, complementary interactions between modalities.

From a technical perspective, multimodal integration offers two critical benefits. It increases overall robustness by adapting its weights: when one modality is noisy or missing (e.g., low-light video conditions at night, background noise in audio), others can compensate, and it enhances interpretability while correlated changes across modalities provide clinicians with stronger evidence of a true event requiring intervention. However, a multimodal model is costly in terms of time and memory. A real-time implementation is not currently feasible within the scope of this project and still needs to be explored (see Section 5.4. Suggestions for future research).

3.5. Model training

3.5.1. Fine tuning and hyperparameters

Both models investigated in this project, MoViNet-A5 and PredFormer, were pre-trained on large-scale datasets before being adapted to the clinical monitoring task. MoViNet was trained on Kinetics-700, a dataset containing approximately 650,000 video clips across 700 human action categories, such as walking, sitting, or interacting with objects. This dataset

provides a broad representation of motion and gesture dynamics, enabling the model to learn generic spatiotemporal patterns. PredFormer, on the other hand, was pre-trained on Human3.6M, a motion capture dataset with millions of frames depicting 3D human poses across diverse activities, including sitting, walking, smoking, and discussion. While Kinetics-700 focuses on high-level action categories from natural videos, Human3.6M emphasizes fine-grained body dynamics, making PredFormer particularly suited for modeling subtle postural changes.

Fine-tuning is the process of adapting these pre-trained models to a specific downstream task by retraining them on a smaller, domain-specific dataset, in this case, simulated hospital videos annotated with intervention labels. The early layers of the networks, which capture low-level features such as edges, textures, and short-term motion patterns, were frozen to preserve their general representational power. Only the later layers, particularly the classification heads, were retrained to specialize in the binary classification task: detecting whether an intervention was required (label '1') or not (label '0'). This strategy significantly reduced the risk of overfitting given the limited dataset size while leveraging the wide knowledge learned from Kinetics-700 and Human3.6M.
From a mathematical perspective, fine-tuning relies on backpropagation, where the gradient of the loss function $\mathcal{L}$ with respect to the model parameters θ is computed using the chain rule:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial y}{\partial \theta}$$

with y the model's prediction. The parameters are then updated iteratively through gradient descent:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta \mathcal{L},$$

where η is the learning rate. Optimizers such as Adam adapt this process by dynamically adjusting the learning rate per parameter based on momentum and past gradient magnitudes, which improves convergence stability.

In this project, several hyperparameters were critical to model performance. The learning rate was set between 1×10^{-4} and 1×10^{-5}, allowing gradual adaptation of weights without overwriting the valuable pre-trained representations. The batch size varied between 4 and 16 depending on GPU memory availability, balancing stability and training efficiency. A dropout rate was applied to the final layers to mitigate overfitting, particularly relevant given the small dataset. The number of training epochs was tuned to ensure convergence without overtraining. Performance was evaluated through k-fold cross-validation, ensuring robustness despite data scarcity.

Fine-tuning is especially pertinent in our case because training a deep video recognition model entirely from scratch would require extremely large datasets, often millions of annotated video clips, and substantial computational resources, which are far beyond what is available in this project. By starting from models pre-trained on large-scale datasets such as Kinetics-700 or Human3.6M, we leverage the fact that these networks have already learned general spatiotemporal features such as motion dynamics, body poses, and common action patterns. These features are transferable to our clinical setting, where the fundamental challenge is still to recognize specific types of human movements and behaviors. Fine-tuning allows us to adapt these broad representations to a more specialized binary classification task (intervention vs. no intervention) using a much smaller, domain-specific dataset. This approach reduces training time, mitigates the risk of overfitting, and ensures that the model can achieve robust performance despite the limited availability of labeled hospital data.

3.5.2. Optimizers, loss functions, and dropout

The training of deep neural networks relies on optimization algorithms to iteratively adjust the model's parameters in order to minimize the loss function. In this project, we experimented with widely used optimizers such as Adam, RMSprop, and Stochastic Gradient Descent (SGD). Adam was ultimately chosen as the primary optimizer because it combines the advantages of both momentum and adaptive learning rates. This makes it particularly efficient when training on relatively small, noisy datasets such as ours, where rapid convergence and robustness to sparse gradients are essential. Adam's update rule dynamically scales the learning rate for each parameter based on estimates of the first and second moments of the gradients, reducing the need for extensive manual tuning.

The loss function defines the training objective. Since the project's task is binary classification, we employed binary cross-entropy loss, which measures the difference between the predicted probabilities and the true labels. Mathematically, for N samples, the binary cross-entropy loss is expressed as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\, log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)\right]$$

where $y_i \in \{0, 1\}$ is the ground-truth label and $\hat{y}_i$ the model's predicted probability of an intervention. This formulation penalizes incorrect predictions more strongly when the model is confident, thereby encouraging calibrated probability estimates.

One of the main challenges when training deep learning models is overfitting. Overfitting occurs when a model learns patterns that are too specific to the training dataset, such as noise or incidental correlations, instead of capturing generalizable features. This results in excellent performance on the training set but poor accuracy on unseen data. Overfitting is particularly problematic in medical applications, where datasets are often limited and the ability to generalize is critical for patient safety.
To mitigate this issue, we applied dropout in the final layers of the network. Dropout works by randomly deactivating a fraction p of neurons during training, which forces the network to learn more distributed and robust representations rather than relying on specific co-adapted features. In practice, we used dropout rates between 0.3 and 0.5 depending on the model architecture. This technique helped improve generalization to unseen data while maintaining training stability.

3.5.3. Frameworks: Tensorflow vs Pytorch

Deep learning has evolved around two dominant frameworks: TensorFlow, developed by Google, and PyTorch, developed by Meta. Although both libraries allow researchers to design, train, and evaluate deep neural networks, they embody different philosophies that influence how models are implemented. TensorFlow was initially designed with production and large-scale deployment in mind, relying on static computation graphs that can be optimized for efficiency but are sometimes difficult to debug. PyTorch, in contrast, was designed around dynamic computation graphs, which make experimentation and debugging considerably more intuitive. réference

In this project, TensorFlow was employed for experiments with MoViNet, because the model is officially provided through the TensorFlow Model Garden together with pretrained weights. This ecosystem allowed for a relatively straightforward fine-tuning process using Keras, while also ensuring compatibility with TensorFlow's deployment tools such as TensorFlow Lite, which are relevant for future real-time applications. Nevertheless, TensorFlow's graph-based nature required a certain adaptation, particularly in controlling

memory usage during video data processing, where large datasets can easily saturate system RAM.

Conversely, PyTorch was necessary for the work with PredFormer, as pretrained implementations based on the Human3.6M dataset are only available in this framework. PyTorch's dynamic execution enabled faster prototyping and clearer debugging when adapting the model to a new classification task. However, PyTorch is less tightly integrated with deployment pipelines, which means that the focus remained primarily on research-oriented fine-tuning rather than production readiness. Switching from TensorFlow to PyTorch also required adapting the data pipeline, since the video clip generators initially written for TensorFlow datasets had to be reformulated into PyTorch's Dataset.

Overall, the dual use of both frameworks highlighted their complementary strengths. TensorFlow provided a stable and well-supported environment for leveraging MoViNet in a way that is directly connected to real-world deployment scenarios, while PyTorch offered the flexibility needed to explore an advanced research model such as PredFormer. This dual approach required additional effort in adapting code and training pipelines, but it provided valuable insights into how each ecosystem supports different stages of the machine learning workflow.

3.6. Evaluation

3.6.1. Evaluation metrics

To rigorously assess the performance of the AI monitoring system, we relied on several standard metrics commonly used in classification tasks: accuracy, precision, recall, and F1-score. Each metric highlights a different aspect of the model's predictive ability, and their combination provides a comprehensive evaluation.

- Accuracy measures the overall proportion of correct predictions and is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

where TP are true positives, TN true negatives, FP false positives, and FN false negatives.
While accuracy offers a general sense of performance, it can be misleading in imbalanced datasets, such as ours, where the majority of time corresponds to "no intervention" (class 0).

- Precision focuses on the reliability of positive predictions and is given by:

$$Precision = \frac{TP}{TP+FP}$$

It answers the question: when the system predicts that an intervention is needed, how often is it correct? High precision is important to reduce unnecessary false alarms that may cause fatigue among medical staff.

- Recall, or sensitivity, measures the ability of the model to detect actual interventions:

$$Recall = \frac{TP}{TP+FN}$$

In our medical context, recall is the most critical metric because false negatives (missed interventions) could result in severe consequences for patient safety. For clinical deployment, recall must approach 100%, as even a single missed event is unacceptable.

- The F1-score provides a balanced metric by combining precision and recall into their harmonic mean:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

This score is particularly useful when evaluating systems under class imbalance, as it penalizes extreme disparities between precision and recall.

In the context of patient monitoring in Intermediate Care Units (IMUs), the metrics must be interpreted through a clinical lens. The primary objective is to achieve a perfect recall (100%) to guarantee that no critical intervention is missed. Precision should remain sufficiently high (≥ 80%) to avoid overwhelming nurses with false alarms, though some level of false positives is tolerable. Accuracy, while less informative in this context, should ideally remain above 90% to demonstrate overall robustness. Finally, the F1-score should reflect a strong balance, with values ideally ≥ 85%, ensuring that both recall and precision are jointly optimized.
To ensure safety, the system will first operate in parallel with nursing staff for validation. Continuous monitoring and retraining will be necessary to maintain these thresholds in real clinical settings. Only once these metrics are consistently met can the system be considered clinically viable.

3.6.2. Cross-validation and robustness checks

Cross-validation is a robust statistical technique designed to evaluate the generalization ability of a machine learning model. In other terms, it evaluates its capacity to perform well on unseen data rather than merely memorizing the training set. The most widely adopted method in literature and the one used in this project, is k-fold cross-validation, where the dataset is split into k equally sized folds. The model is trained k times, each time using $k - 1$ folds for training and the remaining fold for validation. The results are then averaged across all folds to provide an overall performance estimate. Mathematically, if M_i denotes the performance metric obtained on fold i, the cross-validation score is:

$$CV = \frac{1}{k}\sum_{i=1}^{k} M_i$$

This averaging reduces variance in evaluation and mitigates the risks of data bias (caused by specific patient characteristics being overrepresented in a split) and overfitting, particularly relevant in medical contexts where datasets are often small and imbalanced.

In this thesis, cross-validation was critical to assess whether the models (MoViNet and PredFormer) learned generalizable patterns of patient behavior rather than overfitting to specific simulation sessions. By ensuring consistent results across folds, we increased confidence that the models would perform reliably when applied to future real patient datasets.

Beyond raw performance metrics, several signals were monitored during cross-validation to verify that the model was learning appropriately:

- Training vs. validation curves: if the training performance increases while validation stagnates or decreases, this indicates overfitting.
- Stability of metrics across folds: large fluctuations suggest sensitivity to data splits, which would undermine robustness.
- Learning dynamics: observing whether the loss function decreases smoothly without oscillations, which could reflect instability in optimization.

A model that demonstrates stable recall close to 100% across folds, with acceptable precision and without large inter-fold variability, can be considered robust enough for clinical testing. Cross-validation therefore provides both a quantitative and qualitative assurance that the system is not only accurate but also dependable in real-world scenarios.

3.7. Collaboration with the Medical staff

Collaboration with the nursing teams was a cornerstone of this project. Katarina, the head nurse of the Intermediate Care Units at both Solna and Huddinge hospitals, served as the project's anchor point and played a central role in coordinating the clinical side. Together, we organized several meetings and open discussions with nurses from both sites to ensure that the system would be tailored to their clinical realities and seamlessly integrate into their daily workflow. These exchanges were highly valuable: they provided me with a deeper understanding of the nurses' working conditions, while also offering the medical staff clear insights into the technical aspects and deployment strategy of the system.

Since the data collection process relied heavily on the active participation of nurses and assistant nurses, effective communication was vital. To support this, explanatory posters and flyers were designed [Appendix 1], that outlined the project's objectives and practical workflow, ensuring accessibility to all staff members. A "question box" was also installed in the units, giving nurses the opportunity to anonymously submit questions or concerns to the technical team. In addition, small-group meetings were held with particularly engaged nurses who volunteered as key contacts. These nurses contributed directly to resolving practical matters such as camera positioning and the interaction between the monitoring setup and their daily routines.

Figure 14 - Question box in the IMU of Karolinska, Solna

The collaboration was met with strong enthusiasm from the nursing staff, whose motivation and trust were essential to the project's progress. The positive reception even sparked discussions about extending the system to other hospital units, such as neurology care,

highlighting the project's perceived relevance and potential impact. Ultimately, this close and reciprocal collaboration with the nursing teams was essential not only for the success of the simulation-based dataset collection but also for fostering trust, and sustainability of the system within the clinical environment.

4. Results and discussion

4.1. Model training results

4.1.1. MoViNet performance

Training was carried out on a series of five folds, for which metrics were recorded. The following curve shows the evolution of the loss over time, where the coloured lines represent the averages obtained for each dataset, respectively the training and validation data. Within the folds, the loss decreases over time. This means that the model is learning correctly. Furthermore, it is notable that the trend of the curve is the same for both data sets, which shows that the model is not showing any sign of overfitting and is, on the contrary, it suggests a stable learning process.

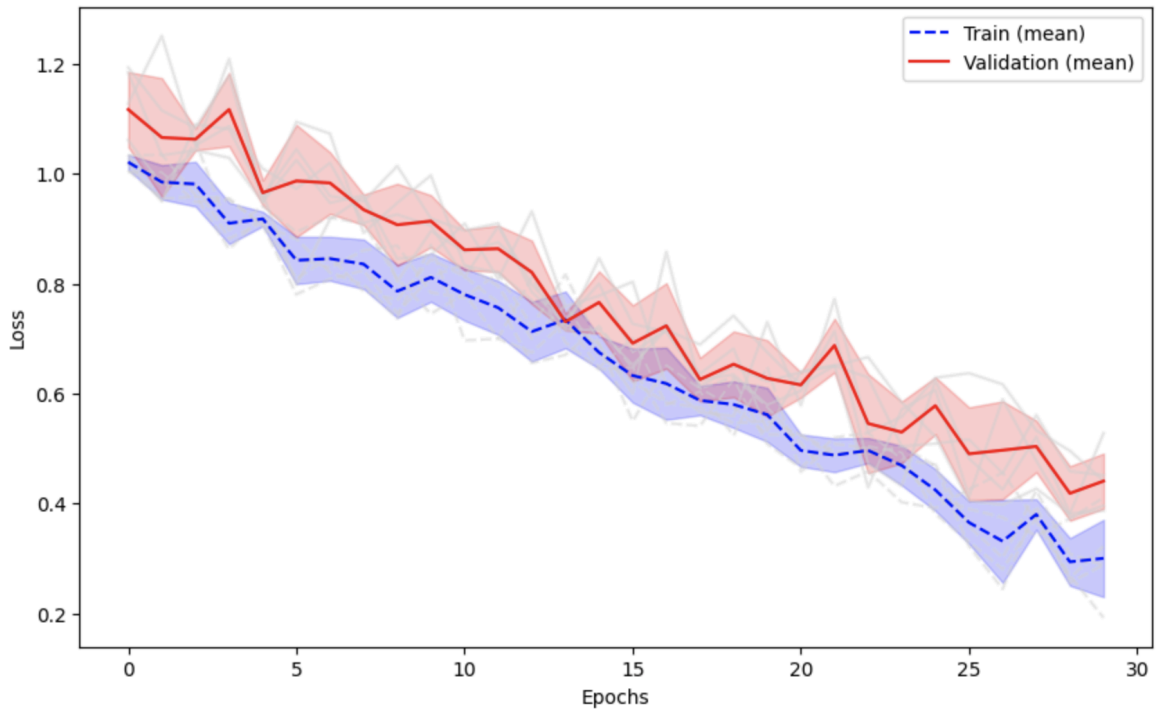


Figure 15 - Training vs validation curves for K-fold cross-validation loss with MoViNet

The confusion matrix for the validation dataset is given below. The notable feature is the absence of false negatives, which are particularly feared in the context of our project since the absence of an alert for a significant event can cause great suffering and have serious

implications. This zero false negative value allows the model to achieve 100% recall, which is one of the key and essential elements of the project.
However, the model occasionally produced false positives, which means that it predicted the need for an intervention when there was not. These are relatively limited but resulted in a slightly lower precision.

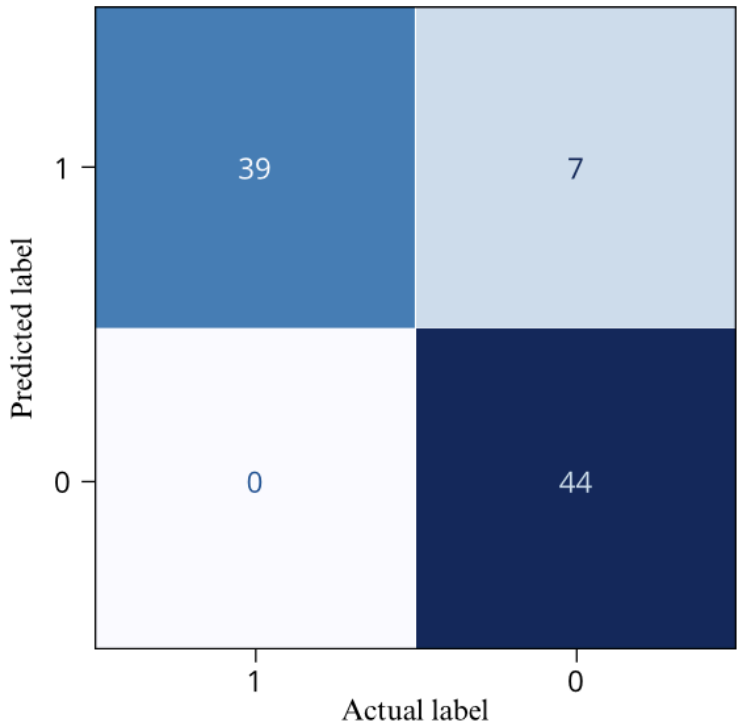


Figure 16 - Confusion matrix of the fine-tuned MoViNet-A5 model using the simulation dataset

This fine-tuned model demonstrated strong performance across all evaluation metrics. It achieved an accuracy of 92%, which confirms that the model is capable of classifying the majority of video clips. The F1-score of 91% confirms the balance achieved between recall and precision.

Metric	Accuracy	Recall	Precision	F1-score
Value	0.92	1.0	0.85	0.91

Table 3 - Results of the fine-tuned MoViNet-A5 model using the simulation dataset

Taken together, these results highlight our fine-tuned MoViNet-A5 model as highly reliable for this task.

4.1.2. PredFormer performance

First and foremost, it is important to note that the batch size used when training the PredFormer model had to be reduced from 16 to 4 due to constantly saturated RAM. Furthermore, training the model took significantly longer than for MoViNet despite this reduction.

As before, training was performed on a set of five folds. The curves showing the evolution of loss over time for each fold, as well as the average of these according to the dataset, are presented below. With the training set, the loss gradually decreases over time, showing relatively stable learning. However, this same curve does not follow the same trend with the validation set. Compared to the previous curves, there are more oscillations and less consistency between folds. This shows signs of overfitting, probably due to the necessary reduction in training data.

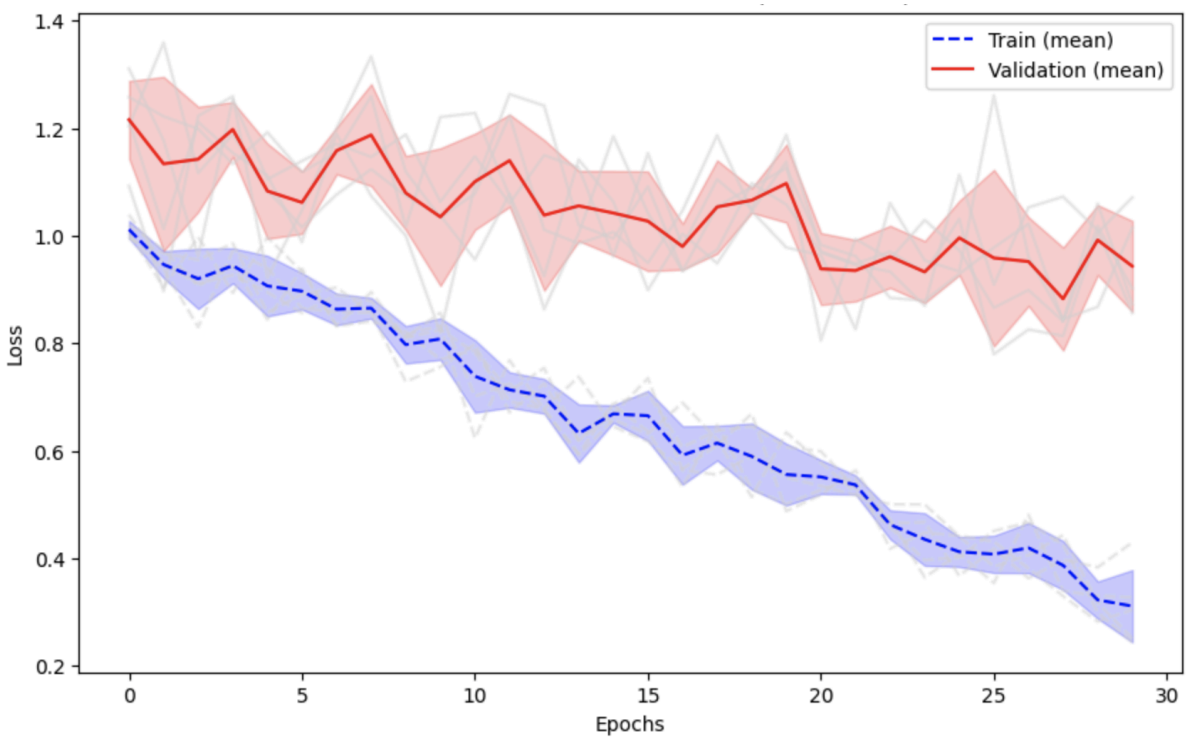


Figure 17 - Training vs validation curves for K-fold cross-validation loss with PredFormer

The confusion matrix for the validation set highlights a number of important points. Firstly, the relatively lower precision is particularly noteworthy, as it indicates a higher rate of false positives and negatives. Three false negatives are counted, which is extremely problematic in our clinical context and causes the recall to drop to 92% compared to 100% with the previous model.

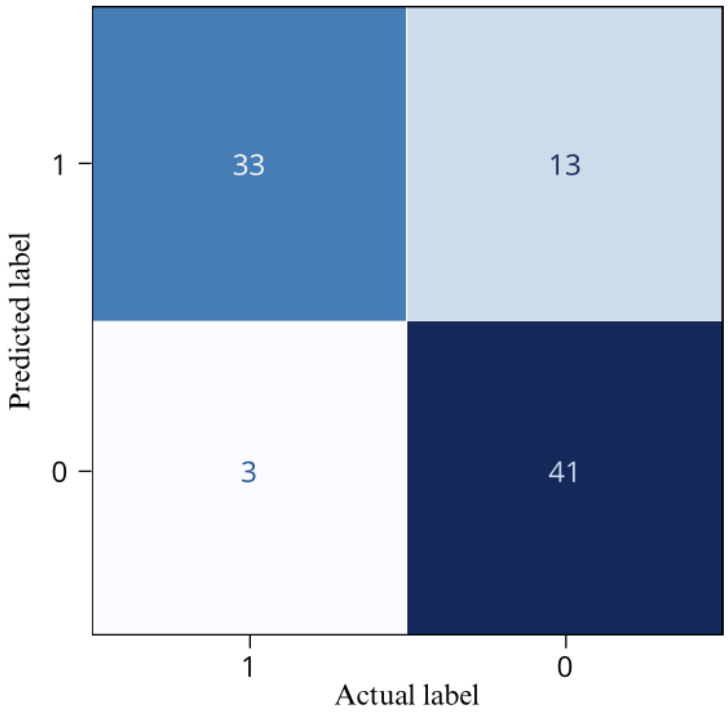


Figure 18 - Confusion matrix of the fine-tuned PredFormer model using the simulation dataset

Metric	Accuracy	Recall	Precision	F1-score
Value	0.82	0.92	0.72	0.81

Table 4 - Results of the fine-tuned PredFormer model using the simulation dataset

The fine-tuned PredFormer model therefore exhibits weaker performance overall. While it demonstrates the capacity to capture relevant events, its computational demands and susceptibility to overfitting with a relatively small dataset limit its practicality in this setting.

4.1.3. Comparative analysis

A direct comparison between MoViNet and PredFormer underscores both the performance gap and the practical trade-offs between the two models. MoViNet not only outperforms PredFormer across all key metrics, but also trains more efficiently, requiring less memory and shorter training times. PredFormer, despite its promising architecture for motion prediction and sequence modeling, appears less suited for the present classification task when operating under constrained computational resources. The need to reduce batch sizes had a negative effect on its ability to generalize, as reflected in the higher rate of false positives and negatives and lower overall precision and recall.

MoViNet's lightweight design and specialized optimization for video classification tasks on the other hand, allowed it to maintain both high recall and balanced precision while training stably across folds.

However, it is important to note that although Kinetics-700 and Human3.6M are both human action recognition datasets, they are different and may also be the cause of the differences observed in the model results.

Nevertheless, these findings suggest that MoViNet is not only the better-performing model but also the more pragmatic choice for deployment in resource-constrained real-world environments as the one this project is aiming for.

4.2. Discussion

4.2.1. Technical constraints

From a technical perspective, several constraints limited the development and evaluation of the monitoring system. A first major limitation was related to the computational resources available for training and inference. Both MoViNet-A5 and PredFormer are computationally expensive models, although to different degrees. While MoViNet is optimized for lightweight inference and therefore compatible with devices such as Raspberry Pis, PredFormer requires significantly more memory and processing power, making it unsuitable for real-time deployment in the current hardware configuration. It is highly likely that the potential of the PredFormer model was limited by RAM and GPU memory restrictions, which led to the decision to reduce the batch size, even though this only serves to prove that the model is not suitable for this particular project.

Another technical constraint involved synchronization and latency in the data pipeline. Although already described as a practical challenge in relation to labeling, it was also a technical issue at the system level. The Raspberry Pi units exhibited delays between the tagging button press and the corresponding frame timestamps, which introduced noise into the dataset. While the implementation of relative frame indexing and CSV-based metadata files reduced this problem, it remains a structural constraint inherent to using low-cost embedded hardware. As this problem was not noticed during laboratory testing, one possible explanation could be the use of a cable that is not shielded from surrounding signals in the hospital room, thereby reducing the flow of transmitted information.

Although the models were pre-trained on large-scale datasets, training on simulation data remains incomplete. Due to organisational difficulties, simulation data could only be collected over a continuous five-hour period, which represents only a tiny fraction of the total data that would be relevant to use. This small dataset contributed to the overfitting of the PredFormer model, despite the augmentation techniques implemented.

Finally, the integration of multimodal data (video, audio, ECG) introduced additional challenges. Each modality required separate preprocessing pipelines, feature extraction, and synchronization before fusion through the multimodal attention mechanism. The impossibility of creating a synchronised and complete dataset meant that this model was only explored as a proof-of-concept, but it paves the way for future optimized experiments.

4.2.2. Practical challenges

Beyond the purely technical aspects, we had to deal with a number of practical and organisational challenges. One of the most significant issues was the lack of patient consent during the project period, which made it impossible to collect real data nor to implement our model in real-time settings, and forced us to use simulation data only. Although this simulated data enabled us to validate the model concept, it does not reproduce the full spectrum of patient behaviours. This gap highlighted the dependency of the project on clinical participation and the complexity of data acquisition in sensitive medical environments.

The project also faced workflow-related challenges. Nurses and assistant nurses already operate under high stress and time pressure, which makes their active involvement in data collection demanding. To address this, communication efforts were made, including explanatory posters, flyers, and the installation of a question box to allow anonymous feedback. These tools proved effective in maintaining engagement, but they also revealed the delicate balance between introducing new technologies into the hospital environment and respecting the limited time and energy of the staff.

Finally, the hospital environment is a challenge in itself. Continuous video recording in patient rooms raises concerns about patient privacy and dignity. Cameras must be turned off when patients are receiving personal care, which requires the creation and design of a simple and intuitive system for turning them off. These considerations demonstrate that, even beyond the algorithms themselves, the design of intelligent algorithms for the healthcare sector requires, above all, the establishment of a strong bond of trust, where respect for privacy and ethics is central.

4.2.3. Ethical and clinical implications

The integration of an AI-based monitoring system into clinical practice raises important ethical and clinical considerations that go beyond the technical development of the model.

A first and central concern relates to patient privacy and dignity. Because the videos collected in this project are not anonymous, they capture patients' faces and expressions, which are essential for detecting subtle behavioral specificities such as pain or distress. While this non-anonymity is necessary for the model's effectiveness, it also places a heavy responsibility on researchers to ensure that data access is strictly limited.

Another ethical implication arises from the subjectivity of interventions. Unlike physiological parameters, which can be quantified through defined thresholds, the decision to intervene often relies on the clinical judgment of nurses, informed by their expertise and experience. This makes it difficult to establish a universally valid ground truth for training. Embedding such subjective judgments into an AI system risks codifying certain practices while neglecting others, which could influence future workflows in unintended ways. Maintaining the system as a tool to support, rather than replace, clinical decision-making is therefore essential. The model must be seen as an assistant that alerts nurses to potential risks, leaving the final decision in the hands of trained professionals.

From a clinical perspective, the implementation of such a system has significant implications for the workload and well-being of medical staff. By reducing the cognitive burden of continuous monitoring, the system has the potential to alleviate stress and fatigue, especially during night shifts where concentration is more difficult to sustain. However, excessive false positives could create additional strain, leading to alarm fatigue and possibly undermining trust in the system. Ensuring that the balance between sensitivity and specificity is clinically acceptable will be key to real-world adoption.

Finally, the project raises broader questions about the responsibility and accountability associated with AI in healthcare. If an alert is missed or misclassified, it remains the clinical staff who must bear the consequences, even though the decision originated from an automated system. This highlights the need for transparent models, interpretable predictions, and clear guidelines on how AI-based alerts should be integrated into clinical workflows. Only by addressing these ethical and clinical dimensions can the system evolve from a research prototype to a trustworthy tool in patient care.

5. Conclusions and future work

5.1. Summary of the work

This thesis explored the development of an AI-based monitoring system designed to assist nurses and assistant nurses in Intermediate Care Units (IMUs) by detecting patient behaviors that may require clinical intervention. Building on prior work and under the framework of an ethical approval, the study investigated two deep learning architectures: MoViNet, a lightweight 3D convolutional network optimized for real-time inference, and PredFormer, a transformer-based model better suited for offline recognition of complex temporal patterns. The project combined technical experimentation with close collaboration with nursing staff, ensuring that the designed system was aligned with clinical needs and workflows. Due to the limited availability of patient data, a simulated dataset was created in a hospital room setting, allowing validation of the technical work and preliminary training of the models. Additional modalities such as audio and ECG were also integrated conceptually through a multimodal attention mechanism but could not be tested due to the lack of synchronized data. Despite the constraints, the study demonstrated the feasibility of implementing such a system in clinical environments and laid the groundwork for future research focused on real-time monitoring, multimodal fusion, and large-scale clinical validation.

5.2. Key findings

One of the key outcomes of this work is the confirmation that lightweight architectures such as MoViNet-A5 can be fine-tuned and adapted for use in a clinical environment, offering real-time inference capabilities suitable for deployment on modest hardware like Raspberry Pi devices. At the same time, the experiments with PredFormer highlighted the potential of transformer-based models to capture long-range temporal dependencies, even though their computational demands currently prevent real-time use. This dual exploration underscored the trade-off between efficiency and accuracy that must be balanced in hospital applications.

Another major finding was the relevance of multimodal integration for robust patient monitoring. While video formed the backbone of the system, the addition of modalities such as audio and ECG signals provided complementary information that could improve the detection of complex or subtle behaviors, such as agitation or early signs of distress. The

attention mechanism, particularly the multi-head design, proved to be an effective way of weighting contributions across modalities, paving the way for more context-aware and reliable systems.

Finally, the project revealed that the success of such AI-based solutions depends not only on technical performance but also on clinical collaboration. Engaging nurses and assistant nurses in the early stages of the system's design ensured that our tool was clinically relevant, and aligned with real-world workflows. Their feedback highlighted the need for a system that minimizes false negatives, tolerates some false positives, and remains intuitive to use under stressful working conditions. This collaborative approach represents a critical step toward ensuring that future iterations of the system are both technically effective and practically sustainable in hospital environments.

5.3. Limitations of the study

As discussed previously, a central limitation of this thesis lies in the lack of real patient data during the development phase. The absence of real patient data makes it impossible to validate the model and system other than through a proof-of-concept.
As the simulation dataset is small in size and diversity, model evaluation is based on only a limited number of examples, which increases the risk of overfitting and limits the generalisability of the results.

Finally, from a technical standpoint, our system also has some limitations. The cameras can only capture images with sufficient lighting, making it impossible to collect data at night. This represents a significant gap, since night shifts are particularly challenging for nurses and may be when automated assistance is most needed. The technical infrastructure also imposed constraints in terms of available resources, which we had to adapt to in order to find an acceptable compromise in the model training results.

5.4. Suggestions for future research

As part of this project, it is necessary to collect a dataset consisting of real patient data. This will not only enable the technical models to be validated using more extensive and complex datasets, but will also strengthen collaboration with medical staff, whose participation in tagging the videos is essential.
Once these models have been validated using real data, real-time implementation can begin. This involves first designing a nurse alert system that is consistent with their way of working, simple and intuitive. Then, the system will be implemented in parallel with the nurses' shifts in order to test the tool's effectiveness and reliability.

Another avenue of research involves creating a synchronised dataset of real patient data, including tagged videos and associated audio, as well as PMS signals. This will enable the multimodal attention model to be validated or rejected, and a decision made on its possible implementation in a hospital setting.

Finally, developing mutual trust between medical and technical staff enables the design of systems that are fully adapted to the everyday reality of nurses. Their advice, opinions, and suggestions must be at the heart of future infrastructure designs so that they are best suited to their needs once real-time implementation begins.

References

[1] Benjamin Jefford-Baker. Autonomous patient monitoring in the intermediate care unit by live video analysis, 2022.

[2] Karolinska University Hospital, "About us," [Online]
https://www.karolinskahospital.com/

[3] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[5] Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. Nature.

[6] Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. NIPS.

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR.

[8] Vaswani, A. et al. (2017). Attention is all you need. NeurIPS.

[9] Tran, D. et al. (2015). Learning spatiotemporal features with 3D convolutional networks. ICCV.

[10] Rajpurkar, P. et al. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. Nature Medicine.

[11] Esteva, A. et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature.

[12] Topol, E. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, vol. 25, 2012.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprintarXiv:1409.1556, 2014.

[15] C.Szegedy, W.Liu, Y.Jia, P.Sermanet, S.Reed, D.Anguelov, D.Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp.1–9.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. InProceedings of the IEEE international conference on computer vision, 2015, pp.4489–4497.

[18] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. InProceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6202–6211.

[19] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong. Movinets: Mobile video networks for efficient video recognition. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16020–16030.

[20] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. InProceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7083–7093.

[21] L.Sun, K.Jia, K.Chen, D.-Y.Yeung, B.E.Shi, and S.Savarese. Lattice long short-term memory for human action recognition. InProceedings of the IEEE international conference on computer vision, 2017, pp.2147–2156.

[22] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. InProceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[23] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang. Tea: Temporal excitation and aggregation for action recognition. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 909–918.

[24] M. S. Ryoo, A. Piergiovanni, M. Tan, and A. Angelova. Assemblenet: Searching for multi-stream neural connectivity in video architectures. arXiv preprintarXiv:1905.13209, 2019.

[25] G.Bertasius, H.Wang, and L.Torresani. Is space-time attention all you need for video understanding. arXiv preprintarXiv:2102.05095, vol.2, no.3,p.4, 2021.

[26] Cournan, Michele DNP, RN, CRRN, ANP-BC, FNP; Fusco-Gessick, Benjamin MA; Wright, Laura RN, CRRN. Improving Patient Safety Through Video Monitoring. Rehabilitation Nursing 43(2):p 111-115, 3/4 2018.

[27] Woltsche, R.; Mullan, L.; Wynter, K.; Rasmussen, B. Preventing Patient Falls Overnight Using Video Monitoring: A Clinical Evaluation. Int. J. Environ. Res. Public Health 2022, 19, 13735.

[28] Sand-Jecklin, Kari EdD, RN, AHN-BC; Johnson, Jennifer BSN, RN, CNRN; Tringhese, Amanda BSN, RN-BC; Daniels, Christine MBA, MSN, RN, NE-BC; White, Freda MSN, MBA, RN. Video Monitoring for Fall Prevention and Patient Safety: Process Evaluation and Improvement. Journal of Nursing Care Quality 34(2):p 145-150, April/June 2019.

[29] S. Yeung, F. Rinaldo, J. Jopling, B. Liu, R. Mehra, N. L. Downing, M. Guo, G. M. Bianconi, A. Alahi, J. Lee et al., "A computer vision system for deep learning-based detection of patient mobilization activities in the icu," NPJdigitalmedicine, vol. 2, no. 1,pp.1–5, 2019.

[30] S. Liu, Y. Yin, and S. Ostadabbas. In-bed pose estimation: Deep learning with shallow dataset. IEEE journal of translational engineering in health and medicine,vol. 7, pp. 1–12, 2019.

[31] F.Bu, Q.Lin, and J.Allebach. Bed exit detection network (bednet) for patients bed-exit monitoring based on color camera images. Electronic Imaging, vol.2021,no. 8, pp.269–1, 2021.

[32] A. Davoudi, K. R. Malhotra, B. Shickel, S. Siegel, S. Williams, M. Ruppert, E. Bihorac, T. Ozrazgat-Baslanti, P. J. Tighe, A. Bihoracetal. Intelligent icu for autonomous patient monitoring using pervasive sensing and deep learning. Scientific reports, vol. 9, no. 1, pp. 1–13, 2019.

[33] Paul Ekman and Wallace V Friesen. Facial action coding system. Environmental Psychology & Nonverbal Behavior, 1978.

[34] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science, 6(2):20, 2016.

[35] Abbe, JacQualine Renee DNP, RN, CMSRN; O'Keeffe, Christian BSN, RN, ONC. Continuous Video Monitoring: Implementation Strategies for Safe Patient Care and Identified Best Practices. Journal of Nursing Care Quality 36(2):p 137-142, April/June 2021.

[36] A. Hajr, B. Tarvirdizadeh, K. Alipour and M. Ghamari, "Monitoring of Four Vital Signs Using Video Processing Based on Machine and Deep Learning Approaches: A Review," 2024 12th RSI International Conference on Robotics and Mechatronics (ICRoM), Tehran, Iran, Islamic Republic of, 2024, pp. 738-744, doi: 10.1109/ICRoM64545.2024.10903541.

[37] Dai, PY., Wu, YC., Sheu, RK. et al. An automated ICU agitation monitoring system for video streaming using deep learning classification. BMC Med Inform Decis Mak 24, 77 (2024).

[38] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. InProceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.

[39] Patrick Lucey, Jeffrey F Cohn, Iain Matthews, Simon Lucey, Sridha Sridharan, Jessica Howlett, and Kenneth M Prkachin. Automatically detecting pain in video through facial action units. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(3):664–674, 2010.

[40] Braeken, A.; Porambage, P.; Gurtov, A.; Ylianttila, M. Secure and Efficient Reactive Video Surveillance for Patient Monitoring. Sensors 2016, 16, 32.

[41] Stefanie Jauk, Diether Kramer, Alexander Avian, Andrea Berghold, Werner Leodolter, and Stefan Schulz. Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. Journal of medical systems, 45:1–8, 2021.

[42] Lindroth, H.; Nalaie, K.; Raghu, R.; Ayala, I.N.; Busch, C.; Bhattacharyya, A.; Moreno Franco, P.; Diedrich, D.A.; Pickering, B.W.; Herasevich, V. Applied Artificial Intelligence in Healthcare: A Review of Computer Vision Technology Application in Hospital Settings. J. Imaging 2024, 10, 81.

[43] J. Hathaliya, P. Sharma, S. Tanwar and R. Gupta. Blockchain-Based Remote Patient Monitoring in Healthcare 4.0. 2019 IEEE 9th International Conference on Advanced Computing (IACC), Tiruchirappalli, India, 2019, pp. 87-91

[44] Jones, Katherine J. PT, PhD; Haynatzki, Gleb PhD, Sabalka, Lucas PhD. Evaluation of Automated Video Monitoring to Decrease the Risk of Unattended Bed Exits in Small Rural Hospitals. Journal of Patient Safety 17(8):p e716-e726, December 2021.

[45] Yujin Tang, Lu Qi, Fei Xie, Xiangtai Li, Chao Ma, Ming-Hsuan Yang, "Video Prediction Transformers without Recurrence or Convolution", arXiv:2410.04733, 2025

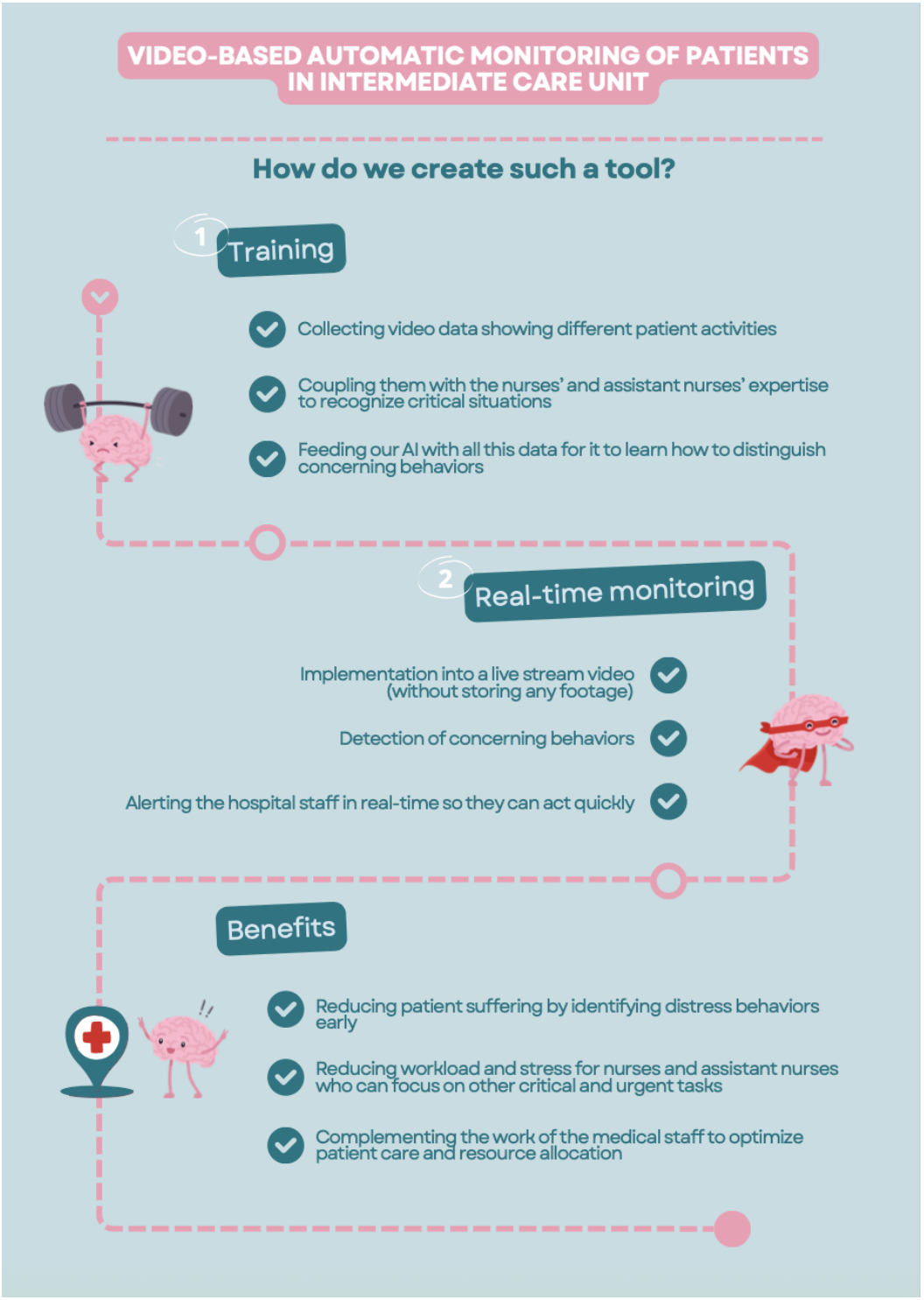
[46] Vista Vascular Clinic, "Brain Aneurysm Treatment — Coiling & Flow," Vista Vascular, Coimbatore. [Online]. https://www.vistavascular.com/areas-of-expertise/brain-aneurysm/

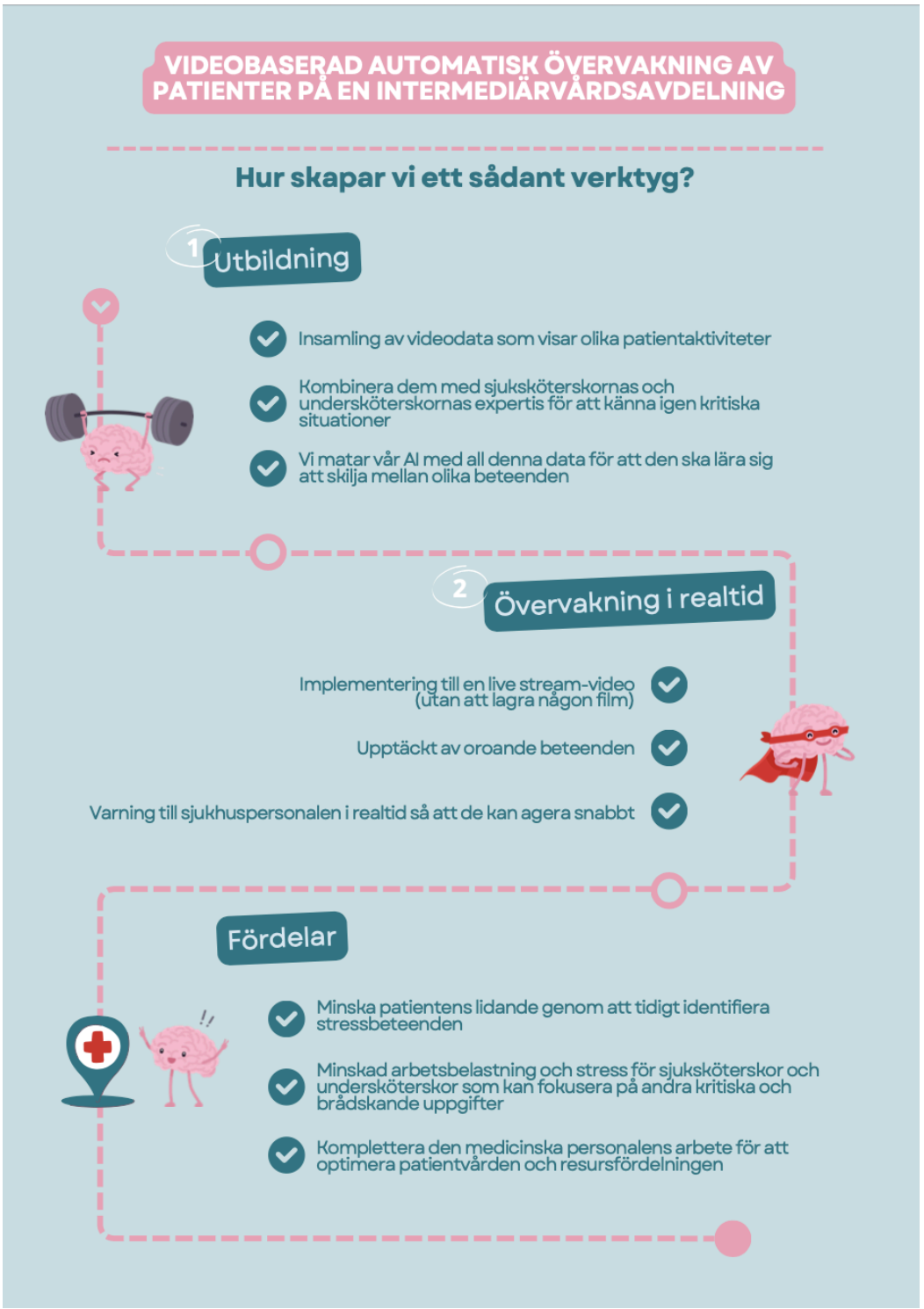
[47] European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). [Online]. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Appendixes

Appendix 1

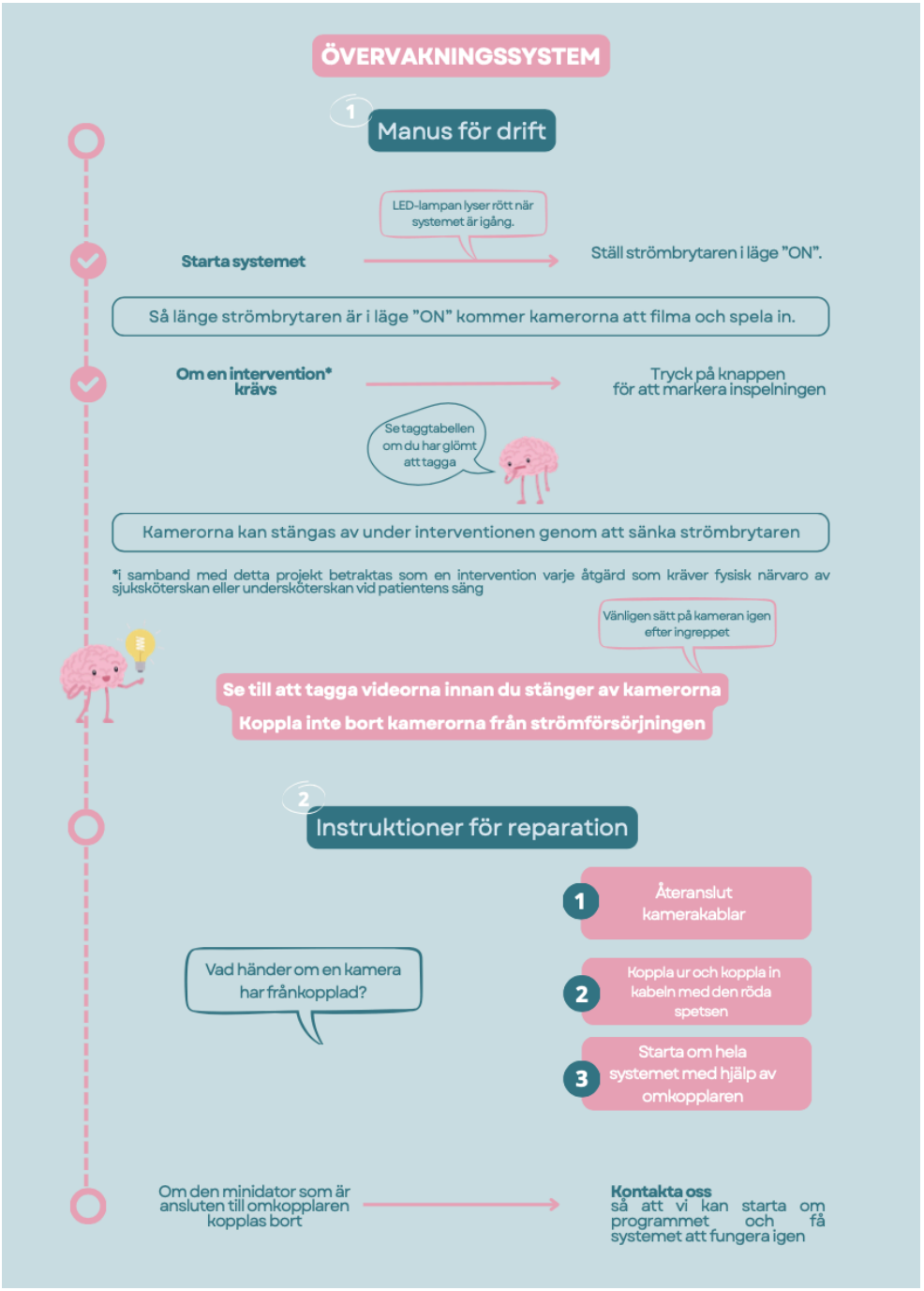
Explicative poster of the project from a technical point of view (English and Swedish versions)




VIDEOBASERAD AUTOMATISK ÖVERVAKNING AV PATIENTER PÅ EN INTERMEDIÄRVÅRDSAVDELNING
Hur skapar vi ett sådant verktyg?
1
Utbildning
Insamling av videodata som visar olika patientaktiviteter
Kombinera dem med sjuksköterskornas och undersköterskornas expertis för att känna igen kritiska situationer
Vi matar vår AI med all denna data för att den ska lära sig att skilja mellan olika beteenden
2
Övervakning i realtid
Implementering till en live stream-video (utan att lagra någon film)
Upptäckt av oroande beteenden
Varning till sjukhuspersonalen i realtid så att de kan agera snabbt
Fördelar
Minska patientens lidande genom att tidigt identifiera stressbeteenden
Minskad arbetsbelastning och stress för sjuksköterskor och undersköterskor som kan fokusera på andra kritiska och brådskande uppgifter
Komplettera den medicinska personalens arbete för att optimera patientvården och resursfördelningen

Appendix 2

Explicative script of the camera and button setup and associated tagbook



TAGGBORD

Om du har glömt att tagga din intervention, vänligen fyll i denna tabell så att vi ändå kan hitta de uppgifter som saknas

Interventioner som du har glömt att tagga	
Datum	**Tidpunkt för interventionen**

Appendix 3
Extract of the Ethical approval and its English translation (generated using DeepL)

Svenska

6.1 Redogör för datainsamling och datas karaktär.
Data från paenter består av video från filmkamera inklusive ljud samt data från paentmonitorerings system (PMS) bestående av EKG, blodtryck, puls, saturaon, temperatur och andningsfrekvens. Data i videon markeras i systemet via e tryck på en knapp/klocka kopplat ll kameran. Parallellt med dea sker en notering i en loggbok av sjuksköterskan eller undersköterskan på vårdrummet.
För a kunna följa upp paenter som vårdats på IMA eer a vården avslutats, kommer en lista med paenters personnummer och vårdd a genomföras. Denna lista är enbart för a vi ska kunna komma i kontakt med paenten och kasseras direkt eer uppföljningen. (Dea ingår inte i forskningen utan är e sä a följa upp paenternas upplevelse av deltagandet i forskningen.)
Data inhämtning via workshops består av ljudinspelning, observaoner och anteckningar. I den första workshopen där sjuksköterskor och undersköterskor llsammans med teknikutvecklare analyserar data bestående av videoklipp. Data från denna workshop innehåller sjuksköterskors, undersköterskor och teknikutvecklarnas analys av videoinspelningar av paenten som ger klassifikaonen för träningen av modellen. I de två följande workshoparna där samutvecklingen av maskininlärningsmodellen sker mellan sjuksköterskor, undersköterskor och teknikutvecklare, kommer ljudinspelning, observaon och anteckningar a genomföras.

6.4 Hur kommer insamlade data a hanteras och förvaras?
Data från video och PMS kommer a bevaras på hårddiskar inom verksamheten på Karolinska Universitetssjukhuset, en i Huddinge och en i Solna. Dessa hårddiskar kommer a vara inlåsta så a enbart deltagare i forskningsprojektet har llgång ll dea, dvs 3 personer (2 teknikutvecklare samt 1 forskare på KTH). Inga paentdata kommer a samlas in utan det är enbart video, dvs kodlistor kommer inte a förekomma i anslutning ll videomaterialet.
Separat kommer K.M. a föra en lista där paentens personnummer framgår, dea för a eer utskrivning från intermediärvårds avdelningen kunna kontakta paenten för e uppföljande samtal.
Data från kalibrerings workshop och de två designworkshoparna kommer a samlas in och handlar om upplevelser, åsikter, och synpunkter ll maskinlösningen och användargränssni. För a säkerställa kvaliteten på forskningen planerar vi a ljudinspela workshopparna samt anteckna.
Ljudinspelningarna kommer a transkriberas. I transkriponerna kommer möjliga namn, platser eller andra idenfierare a tas bort eller ges pseudonymer. Ljudinspelningar och anteckningar från workshoparna kommer a lagras på en lösenordsskyddad, krypterad

hårddisk och låsas in i säkert skåp på Karolinska Universitetssjukhuset. Transkriponer kommer a lagras med en anonym idenfierare (t.ex. P01) i en lösenordsskyddad, GDPR-kompabel molnserver på Karolinska Universitetssjukhuset. Ljudinspelningar kommer a raderas eer a transkriberingen har avslutats. Data utan den personliga idenfieraren kommer a lagras ll 2030. Endast forskarna (se nedan) har llgång ll data.
Enbart forskarna knutna ll dea projekt kommer a ha llgång ll data, dvs 6 personer (forskare från KTH 2 st, Karolinska Instutet 1 st, Karolinska Universitetssjukhuset 1 st och teknikutvecklare 2 st). Materialet kommer a vara anonymiserat och det kommer inte a förekomma några personuppgier i samband med detta material.

English translation (by DeepL)

6.1 Describe the data collection and the nature of the data.
Data from patients consists of video from film camera including sound and data from patient monitoring system (PMS) consisting of ECG, blood pressure, pulse, saturation, temperature
and respiratory rate. The data in the video is marked in the system by pressing a button/clock connected to the camera. In parallel, an entry is made in a logbook by the nurse or the assistant nurse in the care room.
In order to be able to follow up on patients who have been treated at the IMA after the treatment has ended, a list with the patient's social security number and treatment will be implemented. This list is only for us to be able to get in touch with the patient and is discarded directly after the follow-up.
Data collection via workshops consists of audio recording, observations and notes. In the first workshop, nurses and assistant nurses together with technology developers
analyse data consisting of video clips. Data from this workshop includes the analysis of video recordings of the patient by nurses, nursing assistants and technology developers, which provide the classifier for the training of the model. In the next two workshops where the co-development of the machine learning model takes place between nurses, assistant nurses and technology developers, audio recordings, observations and notes will be conducted.

6.4 How will the collected data be handled and stored?
Data from video and PMS will be stored on hard drives within the Karolinska University Hospital, one in Huddinge and one in Solna. These hard drives will be locked so that only participants in the research project have access to them, i.e. 3 people (2 technology developers and 1 researcher at KTH). No patient data will be collected, only video, i.e. code lists will not be present in connection with the video material.

Separately, K.M. will keep a list of the patient's social security number, in order to be able to contact the patient for a follow-up call after discharge from the intermediate care unit.
Data from the calibration workshop and the two design workshops will be collected and is about experiences, opinions, and views of the machine solution and user interface. To ensure the quality of the research, we plan to audio record the workshops and take notes.
The audio recordings will be transcribed. In the transcripts, possible names, locations or other identifiers will be removed or given pseudonyms. Audio recordings and notes from the workshops will be stored on a password-protected, encrypted hard drive and locked in a secure cabinet at Karolinska University Hospital. Transcripts will be stored with an anonymous identifier (e.g. P01) in a password-protected, GDPR-compliant cloud server at Karolinska University Hospital. Transcripts will be stored with an anonymous identifier (e.g. P01) in a password-protected, GDPR-compliant cloud server at Karolinska University Hospital. Audio recordings will be deleted once the transcription has been completed. Data without the personal identifier will be stored until 2030. Only the researchers (see below) have access to the data.
Only the researchers associated with the project will have access to data, i.e. 6 people (researchers from KTH 2, Karolinska Instutet 1, Karolinska University Hospital 1 and technology developers 2). The material will be anonymised and there will be no personal data associated with this material.