Autonomous Medical Simulators

Sapir Gershov

Autonomous Medical Simulators

Research Thesis

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Autonomous Systems
and Robotics

Sapir Gershov

Submitted to the Senate
of the Technion — Israel Institute of Technology
Cheshvan 5782 Haifa October 2021

This research was carried out under the supervision of Dr.Shlomi Laufer, in the faculty of Industrial Engineering and Management.

Some results in this thesis have been published as articles by the author and research collaborators in conferences and journals during the course of the author's master's research period, the most up-to-date versions of which being:

Sapir Gershov, Yaniv Ringel, Erez Dvir, Tzvia Tsirilman, Elad Ben Zvi, Sandra Braun, Aeyal Raz, and Shlomi Laufer. Automatic Speech-Based Checklist for Medical Simulations. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 30–34, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Dr. Shlomi Laufer for the continuous support of my MS.c study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor.
Besides my advisor, I would like to thank Dr. Aeyal Raz, for his encouragement, insightful comments, and hard questions.
My sincere thanks also goes to Dr. Moran Artzi for enlightening me the first glance of research and taught me how to drink Turkish coffee.
Last but not the least, I would like to thank my parents and siblings for supporting me throughout my life and always inspiring me to reach new heights.

The Technion Autonomous Systems Program funding of this research is hereby acknowledged.

Contents

List of Figures

Abstract

In critical and complex life-and-death situations, such as during complex resuscitation procedure, anesthesiologists' decision-making is of utmost importance. Thus, clinical decision support systems (CDSS) have been deployed to assist the medical staff by enhancing clinical decisions. In a field where seconds can make the difference between life and death, integrating an autonomous CDSS framework capable of predicting medical treatment planning and assisting accordingly, can save lives. This requires the framework to accommodate a certain level of awareness and understanding which will not affect clinicians' work except in cases it is required. In addition, the system must withstand diagnostic ambiguity and chaotic environment. In this paper, we describe a technique for mining speech uttered during medical simulations to automatically create plans of resuscitation procedures, which leverage graph networks and language models. Furthermore, during complex resuscitation, we describe a technique for recognising and monitoring medical treatment plans and predicting physician next action. This can be used to save time by prepping the required instrument in advance. This autonomous CDSS can be used to assist anesthesiologists during medical emergencies, and our simulations shows it can save precious minutes in prepping adrenaline dosage, which is crucial for a successful resuscitation.

Abbreviations and Notations

- **ACLS:** Advanced Cardiovascular Life Support.
- **AI:** Artificial Intelligence.
- **CDSS:** Clinical Decision Support System.
- **CNN:** Convolutional Neural Network.
- **GCN:** Graph Convolutional Networks.
- **ICC:** Interclass Correlation Coefficient.
- **KB:** Knowledge Base.
- **KG:** Knowledge Graph.
- **MDP:** Markov Decision Process.
- **ML:** Machine Learning.
- **MLM:** Masked Language Model.
- **NLP:** Natural Language Processing.
- **OR:** Operation Room.
- **RL:** Reinforcement Learning.
- **SPM:** Surgical Process Modelling.

Chapter 1

Introduction

In today's medicine, anesthesiologists play a vital role in the care of physiologically unstable patients, which are often prone to life threatening crises. During medical emergency, medical practitioners are facing diagnostic ambiguity and numerous disruptions in chaotic work environment. To overcome these challenges and provide the best possible medical care, anesthesia professionals must execute highly coordinated strategies in a manner of seconds [1].

The notion of leveraging machine intelligence to assist with difficult decisions during complex medical situations, has fascinated both clinicians and Artificial Intelligence (AI) researchers [2]. The joint efforts of these two domains have produced intelligent technologies that provide situation-specific advice that have an influence on the medical staff decision-making [3]. In the dynamic settings of medical emergency, decision support systems can be valuable tools for optimizing patient care outcomes. The clinical decision-making process is a complex one, since often the medical data and patient information is vague, conflicting, missing, or non-interpretable. Thus, clinical decision support systems (CDSS) require a sophisticated modeling methodology that can handle these challenges and quickly make a decision. CDSS are designed to improve physician's work by enhancing medical decisions with targeted clinical knowledge, patient information, and other medicinal data. With the technical advances in the field of machine learning (ML) in recent years, computers can learn from past experiences and recognize patterns in the clinical data. Therefore, future CDSS frameworks are expected to leverage information and observations otherwise unobtainable or uninterpretable by humans [4]. Moreover, they will become significantly more autonomous, going beyond making suggestions to autonomously performing certain tasks [5].

Contemporary state-of-the-art CDSS are mostly rely on knowledge bases (KBs) generated from expert medical knowledge and clinical treatment guidelines. However, for a CDSS framework to be influential during medical emergencies, it must be based on two fundamental elements: embedded medical knowledge and sequential decision-making. Integrating AI technologies which facilitate these key elements, will provide support to a considerable scope of decisions, particularly in cases of uncertainty. In addition, such systems will gain the ability to manage information from different domains and to evaluate the consequences of the proposed solutions. This is most important when there is an abundance of variables to consider, which makes the decision procedures dramatically more complicated [5].

In this paper we combine embedded medical knowledge with sequential decision-making to

produce a highly intelligent CDSS, which brings the model closer to the ideal CDSS framework compared to the current literature. The key contribution of our research is the construction of a fully automated speech-based framework for real-time anesthesiologist assistance. The developed system is capable of automatic interpretation and reasoning of anesthesia workflow, which allows the system to operate in an independent fashion and to autonomously assist the medical staff in standard and emergency procedures.

The developed system operates as follows: Participants verbal communications were recorded and analyzed using natural language processing (NLP) techniques to preform state estimation. Based on the temporal order of the deduced states, a suitable plan was generated. The constructed plans are the foundations for the system knowledge graph (Figure 1). Next, with advanced methods from language models and graph networks, the developed system uses the collected information, previous and current action, as an input to predict the complete workflow sequence. By constantly accumulating new information and making assumptions based on previous plans, the system performs sequential decision-making to deduce the most accurate plan prediction in any given moment (Figure 2). Finally, the systems state prediction capabilities were evaluated. In addition, we evaluated our CDSS on two benchmarks obtained from medical simulation: medication management and real-time intervention. Our results show we are able to provide timely advice.

Chapter 2

Related Work

2.1 Medical Knowledge Embedding

Knowledge graph (KG), also known as a semantic network, is a directed graph-structured data model used to integrate information of real-world entities (i.e. objects, events, situations) and illustrates the relationship between them. This approach encode knowledge into a form that is amenable to automated analysis and inference. A KG is composed of three key components: nodes, edges, and labels. Any entity can be a node while the edge defines the relationship between the nodes. More formally, given a set of nodes N, and a set of labels L, a KG is a subset of the cross product $N \times L \times N$.

KG embeddings are low-dimensional representations of the entities and relations in a KG. They provide a generalizable context regarding the KG that can be used to infer relations. The KG embeddings are computed so that they follow a given KG embedding model. By using score functions that measure the distance of two entities, we can train and improve KG embedding models. R-GCN [6] model was the first to successfully incorporate graph networks for KG embedding. With the improvement in the methodology of KGs, many researchers have successfully constructed heterogeneous graphs that reflect expert medical knowledge. Medical KGs contain information regarding medicine and clinical practice guidelines, and over the past years they have been constructed from extensive volumes of medical databases. Medical KGs have significant potential for assisting physicians during clinical decision-making in complex scenarios [7].

MedGraph [8] is a graph-based data structure capable of capturing both structural information and temporal sequencing information. However, the proposed method is incapable of performing medical recommendation. SMR [7] is another framework for decomposing medical recommendation into a link prediction process while considering the patient's medical history. This work as well is incompatible in complex clinical scenarios, where time is of great importance.

2.2 Sequential Decision Making

Sequential decisions, a sequence of interrelated decisions over time, are encountered in patient clinical treatment. Due to the uncertainty regarding the effects of a series of treatments over time, the ability to predict clinicians' course of treatment becomes a challenging task. To over-

come these uncertainties, researchers modeled them directly as probabilistic components in a Markov decision processes (MDPs) model [9]. In their paper, Schaefer et. al. presented an efficient method for determining the ideal sequence of decisions\actions in a dynamic and uncertain clinical environment. In addition, the framework allows both qualitative and quantitative analysis of the procedure. However, as the size of the problem increases, MDPs become harder to solve.

Recent studies developed new mathematical and statistical modeling techniques for embedding medical processes into computers. These models pave the way for facilitating clinical decision-making process using neural networks [10].

2.3 Clinical Decision Support Systems

Reasoning with knowledge bases and sequential decision-making are both methods of reasoning in the presence of uncertainty, and have been applied separately in different CDSS frameworks. While both areas have been thoroughly researched and the literature of each is extensive, researchers have yet explored their combined strengths [11]. Recently published literature [12] introduced new method for improving patient care in a data-driven manner, especially in acute care settings. Although helpful, Senders et. al. have not taken into consideration the importance of decisions that are dependent on previous actions. Another paper by Prasad et. al. [13] have modeled process using sequential decision-making, which can be learned by neural networks. In their paper they used reinforcement learning (RL) approach to achieved dynamic and personalized policies in different medical aspects. Nevertheless, the proposed methods are not bound by medical knowledge and guidelines, which is unsettling when safety and accountability is paramount.

Chapter 3

Materials and Methods

Collecting medical emergency data in a hospital is a difficult task due to the unpredictable and complicated nature of the work environment. Furthermore, unlike surgery and other clinical tasks which are typically confined to specific areas in the hospital, advanced resuscitation may accrue sporadically in any location in the hospital. Therefore, in this study we collected data using clinical simulations. Clinical simulation provide a platform for collecting clinical performance data, since they can offer the opportunity to directly observe events in a safe and controlled environment [14]. Furthermore, simulation-based assessment is a method commonly developed for performance-based assessment of medical practitioners [15].

Two simulation scenarios were developed: management of a patient with a severe anaphylaxis reaction and a patient after surgery suffering from severe bradycardia. Both simulations were developed by an experienced anesthesiologist to practice residents for the anesthesiology board certification exam.

Twenty senior anesthesiology residents, 13 males and 7 females, participated in the study. Eight of them performed both simulation scenarios, five residents performed only the anaphylaxis scenario and seven performed only the bradycardia scenario. In addition, two members of the research team played the roles of a nurse and a medical intern. During the simulation, an experienced anesthesiologist evaluated the resident's performance using a checklist. A 'Laerdal' MegaCode Kelly, a full body manikin designed for the practice of Advanced Cardiovascular Life Support (ACLS), was used as the patient (Figure 3.1). The study was approved by the hospital IRB committee.

As part of the study, we developed a simulation setup for data collecting inside the hospital, at the post-surgery recovery unit. Thus, ensuring a realistic environment. The setup included video and audio recording, managed by StreamPix (NorPix Inc.). The recorded video data was collected from 3 different cameras, each recording a different angle of the simulation. In addition, we recorded the patient monitor. As for audio recordings, both the resident and the nurse wore a wireless lavalier microphone transmitter (Sony UWP-D11), which was connected to a digital mixer (Tascam US-20x20). Each audio channel was recorded separately and transcribed afterwards.

Similar to other medical simulation studies [16], a task specific checklist was developed for each scenario. The checklist included approximately 35 tasks the participants were expected to perform. Each task includes short descriptions for the examiners, which guide them in the

Figure 3.1: Image from a simulation.

process of identifying the different assignments performed by the participants.

3.1 Plan monitoring

In our previous research [17], we observed that in most cases the conversation among medical staff may indicate the physical action being performed. By analyzing the participants' speech, we can automatically identify and fill the appropriate rubrics in a task specific checklist. To this end, we developed an end-to-end fully automatic speech-based objective checklist validation system, capable of identifying anesthesia residents' actions based solely on the participants' speech. For each task description the system generates a bag-of-words, a multi-set of the description words, disregarding grammar and word order while preserving multiplicity. The bag-of-words allow us to evaluate how well a sentence in the transcription describes the task in hand. The matching process is based on term frequency–inverse document frequency (TF-IDF) [18] with threshold argmax. This is a numerical statistic that is intended to reflect how important a word is to a sentence in a corpus. According to this approach we multiply two metrics: word frequency in a corpus, and the normalized word frequency where each word count is divided by the number of sentences this word appears in. The higher the score, the more relevant that word is in that particular description. Once we calculated the TF-IDF score of each sentence in the transcription, with respect to the task description in hand, we match the most suitable sentence according to the highest value (threshold argmax). The output of our system is a filled-out checklist, where each task comes alongside with the most suitable sentence in the transcription and the sentence timestamp (Table 3.1). This allows us to verify our system performance.

3.2 From State Estimation to Plan

As mentioned, to fill-in the checklist, our system performs matching process between sentences in the transcription and the task at hand. This can be considered as state estimation. Afterwards, the system produces the marked checklist with the corresponding timestamp for each task. With our previous system output, we can easily transform the filled checklist into a sequence of states,

Task	Timestamp	Evidence in Text
Listening to the lungs and heart	00:03:00.520	”I am listening”
Adrenaline 0.2 - 0.3 mg	00:04:28.040	”Give me Adrenaline. OK, thank you. Adrenaline is inside.”
Ventolin Inhalation - Half a cc of Ventolin	00:03:54.540	”We can do Ventolin inhalation”

Table 3.1: Example of the speech-based checklist system output (translated from Hebrew)

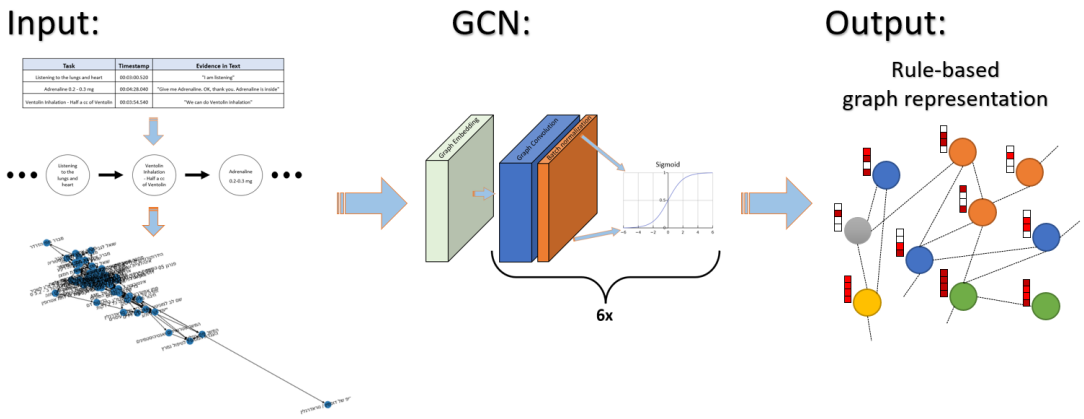

Figure 3.2: General description of the network architecture for embedding medical knowledge and sequential decision-making. **Input:** Marked checklist with the corresponding timestamp for each task is transformed into a sequence of states (plan), in which the nodes represent the performed tasks (state) while the edges represent the order of events. By joining them together we form a directed graph. **GCN:** Graph Convolution network. A series of chronological medical actions are the input of the graph network, of which the model can learn the hidden sequential decision-making and embed medical knowledge. **Output:** The model produces a rule-based graph representation of medical actions nodes.

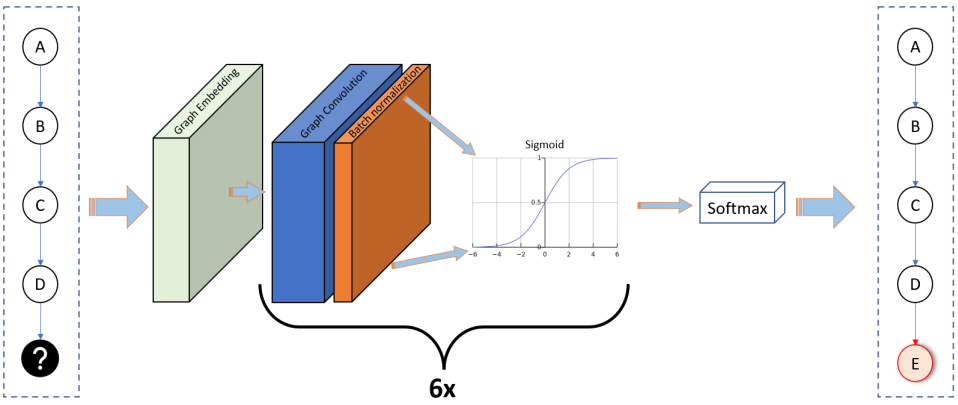

Figure 3.3: General description of step prediction task. The developed system uses the collected information, previous and current action, as an input to predict the next most probable action. By constantly accumulating information, the system performs sequential decision-making to deduce the most accurate prediction in any given moment. In the presented example, the input is the left sequence of letters (A-D) and the black node with the question mark stands for a single step prediction. The right sequence of letters (A-E) stands for the system output, where the letter ’E’ is the predicted step.

and generate a plan in which the nodes represent the performed tasks (state) while the edges represent the order of events. Once we transformed each recorded simulation into a plan, we can join them together to form a graph. This will serve as the infrastructure for the model KG.

3.3 Language Models

The requirement for a model capable of embedding expert medical knowledge, led us to adopt models from the NLP field. With the use of various statistic and probabilistic techniques, language models (LM) encode the complexities of language, the grammatical structures, by learning the features and characteristics of a given language. With a sufficient amount of training data, the model can establish contextual rules. Eventually, for a new corpora data, LM can accurately place distributions over sequence of words and sentences. In traditional LM techniques, during the process of learning grammatical knowledge on vast amounts of data, LMs compactly extract knowledge and information encoded in the training data, and may even store relational knowledge from extensively different contexts [19]. The potential of knowledge transfer from the training data to the model, though limited, is crucial for our task.

Although medical workflow allows shifts in the actions sequentially, the order is not sporadic. Since physician behavior is rule-based by medical knowledge and clinical guidelines, we claim that a computer can find patterns and rules in physicians actions. Just as traditional LM learn grammatical structures from sequences of words, we can apply the same principles on series of chronological ordered medical actions. Due to that, we argue that medical actions relational knowledge is a sufficient substitute for medical knowledge, and therefore, traditional LM has the potential to embed medical knowledge.

Nevertheless, the information presented in our cumulative KG is sparse and often incomplete. This phenomena has devastating results on the process of embedding KGs in LM, and recent publication by He et. al. [20] proposed to harness the potential of graph networks to overcome this challenge. In this study we applied methods of graph networks, specifically the link prediction problem, to find new patterns and sequences in the KG.

To the best of our knowledge, we are the first to use modern LM techniques to embed medical knowledge within a model.

3.4 Graph Convolutional Networks

In the last decade, Convolutional Neural Networks (CNN's) have gained extensive achievements on Euclidean data. However, data in the real world may have underlying graph structures which are non-Euclidean. The non-regularity of data structures has led to recent advancements in new forms of CNNs [21].

Graph Convolutional Networks (GCNs) [22] are an evolved form of CNNs on graphs, which already achieved state-of-the-art results in various application areas [23, 11]. Instead of having an input of 2-D or 3-D arrays, GCN takes a graph as an input. Similar to CNNs, a k-layer GCN is identical to applying a k-layer convolution on the feature vector x_i of each node in the graph. The "graph convolution" applies the same linear transformation to all node's neighbors. The difference is in the hidden representation of each node, as it is normalized with its neighbors

at the beginning of each layer. Afterwards, by stacking layers of filters followed by a nonlinear activation function, the network can learn the graph representations.

The input to graph convolution layer is a set of N node features from embedding layer $\mathbf{h} = \{h_1, h_2, ..., h_N\}$ where $h_i \in \mathbb{R}$ represents the d-dimensional features of i^{th} node; a set of relation types $R = \{r_1, r_2, ..., r_k\}$; and a set of relation features $\mathbf{m} = \{m_1, m_2, ..., m_k\}$, where $m_k \in \mathbb{R}$ is the feature vector of r^{th}-relation type of dimension d.

GCNs have a great expressive power to learn the graph representations and have achieved a superior performance in a wide range of tasks and applications, one of them is the link prediction problem.

3.5 Link Prediction Problem

The link prediction problem is that given the node features X, the model can output whether two nodes are connected by an edge. To be more accurate, in a domain specific graph $G(V, E)$ where $V = \{1, 2, , N\}$ is the node set and $E \subseteq V \times V$ is the link set, GCN utilizes edges $E \in G$ to aggregate and learn node embeddings. The possibility of a connection is decided according to the similarity score of two node embeddings [22]. When it comes to KGs, embeddings are typically used to represent entities and relationships which embody the characteristics of the graph structure. Link prediction uses the resulting vectors to find possible and unobserved associations (links) between two nodes.

In our case, link prediction can find new patterns and associations between different medical action nodes. This allows the model to generate new plans not yet been performed by human participants. In addition, with link prediction we can minimize the difficulty of incomplete KG.

We evaluate the KG embeddings under the closed-world assumption, in which not observed connection between two nodes in a given knowledge graph are false. This assumption transform the evaluation to a well-defined task. Models are judged solely by their ability to fit known data.

In our work we are applying deep learning techniques, which allows us to produce better results even when the input does not obey the graph structure.

3.6 Model Training Objectives

A GCN-based link prediction model optimizes the likelihood of connectivity between two nodes u and v, as a function of the node representation, ${h_u}^{(L)}$ and ${h_v}^{(L)}$, computed from the multi-layer GCN:

$$y_{u,v} = \phi({h_u}^{(L)}, {h_v}^{(L)}) \tag{3.1}$$

Where $y_{u,v}$ is the score between node u and node v. Given an edge connecting u and v, we encourage the $y_{u,v}$ score to be higher than the score between node u and a different node v' from graph G. The Binary Cross-entropy loss function applied in the network training can achieve the behavior above if minimized. More specifically, for each node in G, we generate a node context g_c based on its features. For a given plan input, we mask a node and its edges randomly. During the masking procedure, the graph structure is left unperturbed. Therefore, the training is learnt by maximizing the probability of observing the masked node v_m and its edges e_m based

on the context g_c. According to our model, the whole simulation plans is the graph G, each task/state in the checklist is a node v and the edges reflect the state (actions) temporal order. As for the node context, g_c is generated according to the node neighborhood and the general graph structure.

3.7 Step Prediction

Once the model training objectives achieved satisfying results, i.e, the model learned the chronological order of medical actions, the systems state (action) prediction capabilities were evaluated using step prediction. Step prediction refers to the use of a model to make a prediction ahead in time, where the 'prediction horizon' defines the extent of future prediction [24].

Due to the chaotic nature and noise in our database, developing a model that accurately predict the next action with horizon of one and two steps is a challenging task. The approach we apply for step prediction is as follows: each prediction from a one-step-ahead prediction model is the input for future prediction horizon, and the total loss is composed of the difference between the predicted action and the ground truth in each time step.

For the construction of the plan, we designed an iterative plan generation process and formulated it as a general decision process $M = (S, A, P)$, where $S = \{s_i\}$ is the set of states that consists of all possible intermediate and final plans, $A = \{a_i\}$ is the set of actions that describe the modification made to current plan at each time step, P is the transition dynamics that specifies the possible outcomes of carrying out an action, $p(s_{t+1}|s_t, ...s_0, a_t)$. The procedure to generate a plan can then be described by a sub-plan $(s_0, a_0, ..., s_n, a_n)$, where s_n is the final generated plan. The addition of a sub-plan at each time step can be viewed as a state transition distribution: $p(s_{t+1}|s_t, ..., s_0) = \sum_{a_t} p(a_t|s_t, ...s_0)p(s_{t+1}|s_t, ...s_0, a_t)$, where $p(a_t|s_t, ...s_0)$ is the policy network ϕ_θ. We design a plan generation procedure that can be formulated as a MDP, which requires to satisfy $p(s_{t+1}|s_t, ...s_0) = p(s_{t+1}|s_t)$. Under this property, the policy network only needs the intermediate graph state s_t to derive an action. The action is used by the environment to update the intermediate plan being generated.

3.8 Implementation Details

Computing infrastructure used for running experiments included a single NVIDIA GeForce GTX 1070 GPU with 8 GB of memory, and Linux 20.04.2 LTS operating system. The proposed network was implemented using PyTorch 1.9 [25]. The dimension of contextual node embeddings is set to 1024. The network was trained from scratch for 120 epochs, while the fine-tuning steps were trained with an additional 30 epochs. We also experimented with other settings and found that small changes did not change the results much. Both training and fine-tuning stages were trained with a batch size of 1 using the binary cross-entropy with logits loss function. The model parameters were trained with ADAM optimizer [26] with a learning rate of 0.0001 and 0.001 for training and fine-tuning steps, respectively. The best model parameters were selected based on the development set. In the construction of our network architecture, we applied methods from 'NetworkX' [27] - a Python language package for exploration and analysis of networks and network algorithms.

As mentioned earlier, for each simulation report we generate a chronologically ordered sequence of medical actions in the form of a table. We than transform the table in to a plan using 'NetworkX' and by joining all of the plans together we form a graph. This is the foundation for the framework KG. The embedding of medical knowledge was achieved by GCNs network for graph representation and Link Prediction for finding new connections and patterns in the directed graph. The network architecture is as follows: first, the input graph is passed into a graph embedding layer. For each entity in the graph, we concatenate the various textual attributes to obtain their embeddings. These embeddings form an initial feature vector of entities to be used in the training. Afterwards, we apply 6 layers of graph convolutions and batch normalization followed by the Sigmoid function. The loss is calculated using binary cross-entropy with logits on the link prediction results. Afterwards, once the model was sufficiently trained, we tested the embedded knowledge by predicting the resident actions. The process of predicting the resident next action was based on step prediction with MSE loss function. Using this method, we mask several nodes and edges in a given sequence and the model is expected to predict those masked tasks based on presented sequence. Such training scheme makes this model bidirectional in nature. 3-fold cross validation approach was applied on all results.

3.9 Empirical Evaluation

The usefulness of the system was evaluated using two potential applications:

1. **Medication Management.** Adrenaline injection plays a key role in the successful management of resuscitation. To evaluate the network ability to oversee the resident's medication management, we performed a two stages evaluation: keyword recognition and state prediction.

   (a) Keyword Recognition. We compared the number of times the resident asked for adrenaline and the number accumulated by the system. To evaluate the results we used the interclass correlation coefficient (ICC) [28] - a descriptive statistic that assess the consistency of the quantitative measurements made by different observers measuring the same quantity. ICC score lower than 0.5 is considered the be a poor reliability, 0.5 – 0.75 is considered a moderate reliability, 0.75 – 0.9 is good reliability and greater than 0.9 is excellent reliability. We calculated the ICC score based on One-way random effects formula:

   $$\frac{MS_R - MS_E}{MS_R} \tag{3.2}$$

   Where MS_R is the mean square for rows and MS_E is the mean square for error.

   (b) State Prediction. We designed the system to predict the timing of the adrenaline injection based on state history. The system input is the previous actions performed by the resident, and the output is the next action prediction. We examined the time delta between the resident and the system output.

2. **Defibrillator Management.** As mentioned, two clinical scenarios were deployed during the data collecting: severe anaphylaxis reaction and bradycardia. In the case of brady-

cardia, part of the expected treatment is to use the defibrillator while in the anaphylaxis scenario it is prohibited. As for our system, the task in hand was step prediction – the ability to predict the physician next state based on his previews states. To validate our results, we evaluated two parameters:

(a) Prediction of defibrillator usage – the system should never predict defibrillator in the anaphylaxis scenario and should predict it at least once for the severe bradycardia.

(b) Time reduction – when the clinical scenario requires a defibrillator, quick delivery of the electrical shock may play a critical role in a successful resuscitation. However, bringing and operating the defibrillator may take time. Therefore, we examined how much time before the physician requires the defibrillator, our system predicts it will be needed. We compared this to the average time it took to bring the defibrillator. This represents the potential time the system may save in the delivery of the electrical shock.

Chapter 4

Results

4.1 Model Evaluation

Link prediction results of the entire plan sequence were based on train and validation loss as well as validation and test accuracy. After 115 epochs, our model achieved the following results:

Task	Validation Acc.	Test Acc.
One step prediction	1.0000	0.8947
Two step prediction	0.7202	0.7037

Table 4.1: Model evaluation results

4.2 Medication Management

The system achieved an ICC score of 0.783. This means that the system is capable of correctly identifying and documenting the request for adrenaline injection. Which implies it can identify and follow orders, even in a real environment. In addition, the system was able to predict the adrenaline injection timing faster than the human counterpart. The average time delta was 00:01:28 (1 minute and 28 seconds) in favor of our system.

The number of times the system would have asked for the adrenaline injection, based on the resident simulation plan was evaluated. The predicted results were compared with the expected result using the ICC score. The system has outmatched the resident with ICC of 0.833 while the human participants achieved an ICC score of 0.583.

4.3 Defibrillator Management

In 2 of the 15 bradycardia cases the participant failed to request the defibrillator while the system successfully predicted the need for defibrillator in all the cases. This shows the potential of our system to serve as a decision-support system in a real clinical environment. Out of 13 anaphylaxis simulations, in just one simulation our system failed to predict the correct treatment and suggested the use of a defibrillator. Based on these results, our system is capable of correctly predict the use/misuse of the defibrillator.

We define the time interval of defibrillator usage as follows: the moment the resident instructed another medical personnel to bring in the defibrillator until the moment the manikin recorded the electrical shock. Based on the collected data, the average defibrillator usage time interval was 00:01:26±00:00:46. Based solely on the severe bradycardia simulations, one step prediction has a time gap of 00:00:50±00:00:50 in favor of our system, while the two steps prediction has a time gap of 00:01:12±00:01:18. Thus, the system can shorten the delay in providing the electrical shock.

Chapter 5

Conclusion & Open Questions

In this study we examined the potential of a fully aware CDSS in the context of critical, clinical decision-making. This is considered to be a difficult objective on account of the required level of awareness and understanding expected of the system, which will not constrain clinicians' work except when it needs to. In addition, during medical emergencies, CDSS are expected to encounter ambiguous information and numerous disruptions from the work environments. These factors combined may affect the CDSS performances. Therefore, we constructed a framework for clinical data collecting from anesthesia residents' medical simulations. Since real clinical cases are sporadic and challenging to capture, we used clinical simulation as a suitable alternative. For each collected simulation data, we generated the resident plan and combined them together to establish the system KB. Afterwards, the system can use new collected information, previous and current action, as an input to predict the complete workflow sequence. The system is completely autonomous and a fully automatic pipeline from raw audio files to a complete process plan was established. Since the system only requires audio signals, it doesn't interfere with the medical staff workflow and minimizes the invasion of privacy.

We first tested our systems ability to predict the next step in the procedure. We then evaluated the system applicability using two practical applications: medication management and real-time intervention. Both evaluations have shown promising results on the collected simulations data and therefore, our framework has proven its potential to supervise over the activity during a medical emergency and to assist in a complex decision-making situation.

5.1 Identification and Validation of Actions

For each simulation recording automatic transcription was performed, and afterwards keywords were identified in each sentence. Using these keywords, a matching process between the checklist tasks and the corpus sentences was implemented. The outcome of the algorithm was a filled checklist in which the completed tasks are provided with a matching sentence and timestamp.

The native language of the current participants of this study is Hebrew. This poses a unique challenge common to Morphologically Rich Language. As clearly evident from the results, using lexical analysis improved our system performances, and might have a greater impact on a more complex models. We plan to expand our work to other languages in the future and assess the system performance.

The system was successful in correctly identifying most of the tasks performed by the participants. Yet, one limitation of the system is that it is currently based on keyword matching and not on a more complex model of the conversation. The method in use has limited accuracy, and in addition, only provides a binary score indication whether the task was preformed or not. For example, the current system may indicate a drug was provided but it will not assess the dosage. In order to develop a more complex algorithm, a significantly larger data base is required. We are continuously collecting data that focuses both on a larger number of participants as well as a wide range of clinical scenarios. This will expedite the development of more complex algorithms.

5.2 Prediction of Trajectory

For each collected simulation data, we generated the resident trajectory and combined them together to establish the system KB. Afterwards, the system can use new collected information, previous and current action, as an input to predict the complete workflow sequence.

We first tested our systems ability to predict the next step in the procedure. We then evaluated the system applicability using two practical applications: medication management and real-time intervention. Both evaluations have shown promising results on the collected simulations data and as of that, our framework has proven its potential to supervise over the activity during a medical emergency and to assist in a complex decision-making situation. Yet, the method we used for state estimation has limited accuracy, and in addition, only provides a binary score indication whether the task was preformed or not. In order to develop a more complex algorithm, a significantly larger data base is required. Another limitation we need to consider is the lack of sufficient participants. By exposing the model to a wider range of physicians' behavior we can improve its predictions and certainty. We are continuously collecting data that focuses both on a larger number of participants and a wider range of medical scenarios. This will facilitate the development of more complex frameworks, which will enhance the its functionality.

5.3 Future Research

In the future, the proposed model can be improved by enriching the database representation. Medication dosage, quality of procedure and temporal information can provide the model additional information to deduce better predictions. Another aspect to consider is how these systems can be applied in the hospital and assist the medical staff work without disturbing the workflow.

5.4 Open Questions

Participants behavior. Participants will always approach a simulator differently compared to real life. Two common changes in attitude can occur: (a) hypervigilance, which causes excessive concern because one knows an event is about to occur; and (b) cavalier behaviour, which occurs because it is clear no human life is at stake. These effects may co-exist and counterbalance and may have a dramatic effect on our system performance. This raises the question on rather we can anticipate this behavior or at least minimize its affect.

Unpredicted behavior In rear occasions, a participant may treat a patient in whole new approach that have never been learned in our system. The question here is how to design a network that can comprehend fuzzy logic and unknown behavior.

Appendix A

Appendix

A.1 Network Training

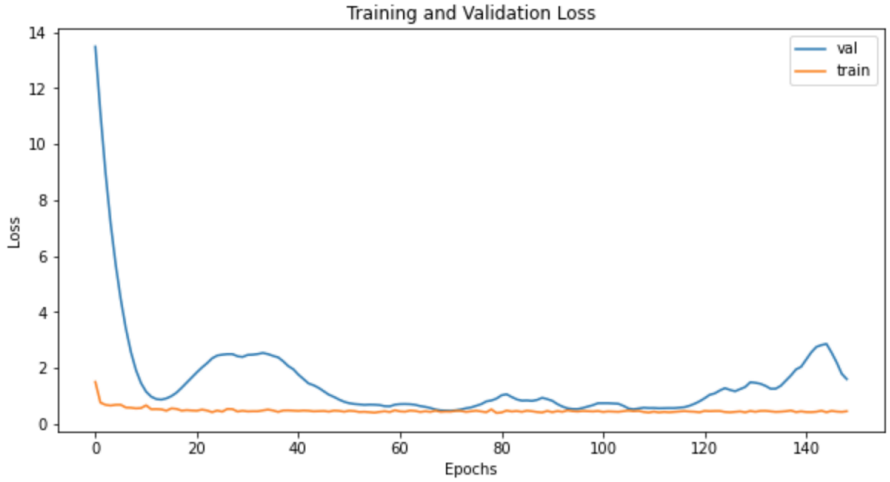

Figure A.1: Network training process. Train vs. Validation Loss Graph.

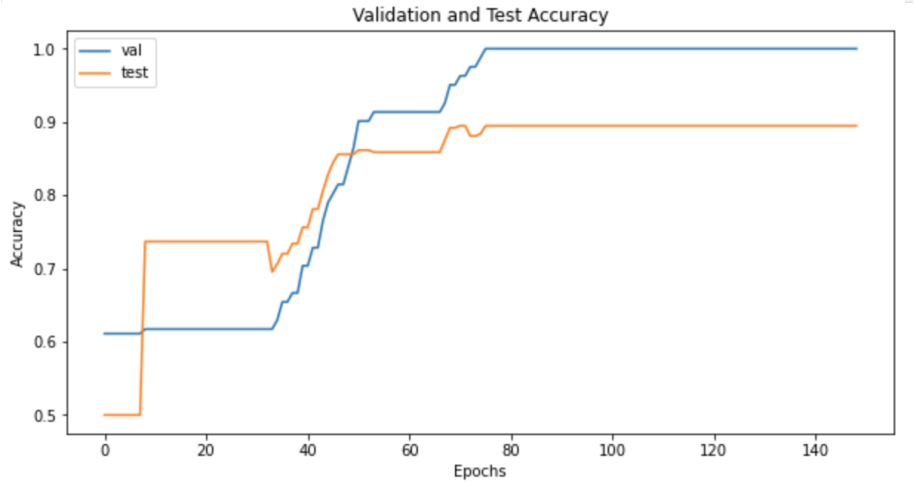

Figure A.2: Network training process. Validation vs. Test Accuracy Graph.

A.2 Network Validation

File Name	Start CPR	End CPR	Delta	Expected	Human Checklist	Transcriprtion
01-07-21_06-56-23.000	00:05:52	00:10:55	0:05:03	2	2	2
01-12-21_06-55-47.000	00:06:02	00:11:18	0:05:16	2	4	4
01-19-21_07-01-24.000	00:06:11	00:09:07	0:02:56	1	3	2
01-21-21_07-13-14.000	00:10:17	00:14:55	0:04:38	2	3	4
01-26-21_06-57-16.000	00:06:22	00:15:47	0:09:25	3	0	0
02-02-21_06-59-20.000	00:04:25	00:09:46	0:05:21	3	3	3
02-25-21_06-48-22.000	00:05:13	00:08:33	0:03:20	2	1	1
03-02-21_07-14-22.000	00:03:35	00:09:33	0:05:58	2	1	1
03-03-21_07-43-17.000	00:07:51	00:10:11	0:02:20	1	1	1
03-03-21_08-15-57.000					0	0
03-03-21_08-43-18.000	00:08:56	00:14:12	0:05:16	2	2	3
03-03-21_09-39-02.000	00:05:18	00:11:10	0:05:52	2	2	2
03-03-21_10-13-28.000	00:06:31	00:10:11	0:03:40	2	2	2
03-03-21_10-57-03.000	00:03:36	00:11:36	0:08:00	3	1	1
03-03-21_11-32-13.000	00:08:30	00:11:34	0:03:04	1	2	2
03-03-21_12-13-49.000	00:06:38	00:10:42	0:04:04	2	2	2
03-03-21_13-47-27.000	00:03:16	00:07:26	0:04:10	2	2	1
03-04-21_06-59-24.000	00:05:44	00:11:48	0:06:04	2	2	2
03-09-21_07-05-10.000	00:07:23	00:09:38	0:02:15	1	2	4
03-11-21_06-54-56.000	00:07:33	00:10:38	0:03:05	1	2	0
03-16-21_07-01-22.000	00:07:40	00:14:11	0:06:31	3	1	1
03-25-21_07-14-46.000	00:04:01	00:11:45	0:07:44	2	2	1
04-27-21_07-16-48.000	00:04:55	00:09:23	0:04:28	2	2	1
04-29-21_06-48-39.000	00:05:36	00:10:40	0:05:04	2	4	3
05-04-21_06-46-03.000	00:04:19	00:11:52	0:07:33	3	0	0
05-18-21_07-18-26.000	00:01:06	00:08:53	0:07:47	3	3	3

Figure A.3: Medication management. Adrenaline keyword recognition.

Machine vs. Human						
File Name	Human First Dosage	Machine First Dosage	Timming	Delta	Average	STD
01-07-21_06-56-23.000	0:07:45	0:07:05	Before	0:00:40		
01-12-21_06-55-47.000	0:05:45	0:05:36	Before	0:00:09		
01-21-21_07-13-14.000	0:09:16	0:07:45	Before	0:01:31		
02-25-21_06-48-22.000	0:06:21	0:05:27	Before	0:00:54		
03-02-21_07-14-22.000	0:07:20	0:03:51	Before	0:03:29		
03-03-21_07-43-17.000	0:08:35	0:06:55	Before	0:01:40		
05-18-21_07-18-26.000	0:04:19	0:01:06	Before	0:03:13		
03-03-21_12-13-49.000	0:07:19	0:07:07	Before	0:00:12	0:01:28	0:01:15
03-03-21_13-47-27.000	0:03:56	0:03:17	Before	0:00:39		
03-04-21_06-59-24.000	0:07:47	0:05:44	Before	0:02:03		
03-09-21_07-05-10.000	0:07:46	0:07:23	Before	0:00:23		
03-11-21_06-54-56.000	0:06:18	0:06:12	Before	0:00:06		
03-16-21_07-01-22.000	0:12:36	0:08:30	Before	0:04:06		
03-25-21_07-14-46.000	0:09:44	0:07:43	Before	0:02:01		
04-27-21_07-16-48.000	0:07:06	0:06:18	Before	0:00:48		
03-03-21_08-43-18.000	0:08:15	0:08:15	On time	0:00:00		
03-03-21_08-15-57.000			Never			
05-04-21_06-46-03.000	Inf	0:05:06	Before			
01-26-21_06-57-16.000	Inf	0:09:39	Before			
01-19-21_07-01-24.000	0:05:45	0:06:33	After	0:00:48		
04-29-21_06-48-39.000	0:05:39	0:06:04	After	0:00:25		
03-03-21_09-39-02.000	0:03:47	0:04:12	After	0:00:25		
03-03-21_10-13-28.000	0:06:13	0:06:35	After	0:00:22	0:00:35	0:00:14
03-03-21_10-57-03.000	0:03:05	0:03:57	After	0:00:52		
03-03-21_11-32-13.000	0:07:33	0:08:26	After	0:00:53		
02-02-21_06-59-20.000	0:04:47	0:05:06	After	0:00:19		

Figure A.4: Medication management. Adrenaline state prediction.

Prediction	Confidance	Timming	Current Action	Current Time	Accurate Timmimg	Duration			
	0.8947	Before	Start Chest Compression	00:03:10	0:05:33	0:02:23			
	0.8354	Before	Adrenaline 1 mg	00:07:47	00:09:22	0:01:35			
	0.7938	Before	Call for help	00:02:20	0:03:00	0:00:40			
	0.8732	Before	Identify VF	00:04:19	0:06:26	0:02:07			
	0.8753	On time	Start Air bag ventilation	00:03:45	0:03:50	0:00:05			
	0.8752	On time	Call for help	00:07:00	0:07:16	0:00:16		Average	STD
One Step prediction	0.9002	On time	Start chest compressions	00:04:38	0:04:51	0:00:13	Time Saved	0:00:50	0:00:50
	0.88	On time	Identify VF	00:04:01	0:04:07	0:00:06	Time Lost	0:00:16	0:00:39
	0.9	On time	Atropine 0.1 - 1 mg	00:02:04	0:02:46	0:00:42			
	0.9004	On time	Atropine 0.1 - 1 mg	0:01:30	0:01:43	0:00:13			
	0.8398	Late	Start chest compressions	0:04:20	0:04:03	0:00:17			
	0.8947	Late	Start Chest Compression	0:07:07	0:06:53	0:00:14			
	0.892	Never	Atropine 0.1 - 1 mg # 3	00:03:52	Inf	#VALUE!	זיהוי טעות של נבחן		Anaphylaxis
									VF
	0.4281	Never	Adrenaline 1 mg	00:08:35	Inf	#VALUE!			
	0.1257	Never	Adrenaline 1 mg	00:08:15	Inf	#VALUE!			
	0.0958	Never	Start chest compressions	00:06:11	Inf	#VALUE!			
	0	Never							
	0.1545	Never	Start chest compressions	00:07:17	Inf	#VALUE!			
	0	Never							
One Step prediction	0.4459	Never	Call for help	00:06:22	Inf	#VALUE!			
	0.2249	Never	Adrenaline 1 mg	00:08:38	Inf	#VALUE!			
	0.2198	Never	Switching roles	00:07:36	Inf	#VALUE!			
	0.4621	Never	Bag valve ventialtion	00:07:45	Inf	#VALUE!			
	0.0568	Never	Start CPR	00:04:25	Inf	#VALUE!			
	0.4787	Never	Start Chest Compression	00:08:23	Inf	#VALUE!			
	0.4784	Never	Call for help	00:06:18	Inf	#VALUE!			

Prediction	Confidance	Timming	Current Action	Current Time	Accurate Timmimg	Duration			
	0.7202	Before	Call for help	00:01:13	0:05:33	0:04:20			
	0.7567	Before	Identify VF	0:03:35	0:04:03	0:00:28			
	0.7562	On time	Call for help	00:03:40	0:03:50	0:00:10			
	0.7812	On time	Connect patient to pacer	00:04:00	0:06:53	0:02:53			
	0.8752	On time	Call for help	00:07:00	0:07:16	0:00:16			
	0.9002	On time	Start chest compressions	00:04:38	0:04:51	0:00:13			
	0.7664	On time	Adrenaline 1 mg	00:08:56	00:09:22	0:00:26		Average	STD
	0.8337	On time	Atropine 0.1 - 1 mg	00:02:53	0:04:07	0:01:14	Time Saved	0:01:12	0:01:18
	0.8024	On time	Start CPR	00:05:06	0:06:26	0:01:20	Time Lost	0:00:41	0:00:47
	0.8125	On time	Start CPR	00:01:06	0:01:43	0:00:37			
	0.7009	Late	Atropine 0.1 - 1 mg	00:02:04	0:02:46	0:00:42			
	0.7009	Late	Call for help	00:02:20	0:03:00	0:00:40			
Two Steps prediction	0.7962	Never	Atropine 0.1 - 1 mg #2	00:02:48	Inf	#VALUE!	זיהוי טעות של נבחן		Anaphylaxis
									VF
	0.4421	Never	Call for help	00:07:51	Inf	#VALUE!			
	0.2238	Never	Call for help	00:08:25	Inf	#VALUE!			
	0.0095	Never	Connect Ambo to Oxygen	00:06:20	Inf	#VALUE!			
	0.4013	Never	Adrenaline 1 mg	00:07:33	Inf	#VALUE!			
	0	Never							
	0.5695	Never	Call for help	00:06:22	Inf	#VALUE!	טעות של המערכת		
Two Steps prediction	0.4377	Never	Adrenaline 1 mg	00:08:38	Inf	#VALUE!			
	0.4753	Never	Switching roles	00:07:36	Inf	#VALUE!			
	0.4621	Never	Bag valve ventialtion	00:07:45	Inf	#VALUE!			
	0.3885	Never	Call for help	00:05:17	Inf	#VALUE!			
	0	Never							
	0.2265	Never	Start CPR	00:07:17	Inf	#VALUE!			
	0.3894	Never	Adrenaline 1 mg	00:06:25	Inf	#VALUE!			

Figure A.5: Real-time intervention. Defibrillator state prediction.

Bibliography

[1] D. M. Gaba, S. K. Howard, K. J. Fish, B. E. Smith, and Y. A. Sowb, "Simulation-based training in anesthesia crisis resource management (ACRM): A decade of experience," *Crew Resource Management: Critical Essays*, vol. 32, no. 2, pp. 349–367, 2017.

[2] Q. Yang, A. Steinfeld, and J. Zimmerman, "Unremarkable AI: Fiting intelligent decision support into critical, clinical decision-making processes," in *Conference on Human Factors in Computing Systems - Proceedings*, vol. 11, ACM, 2019.

[3] V. L. Patel, E. H. Shortliffe, M. Stefanelli, P. Szolovits, M. R. Berthold, R. Bellazzi, and A. Abu-Hanna, "The coming of age of artificial intelligence in medicine," *Artificial Intelligence in Medicine*, vol. 46, no. 1, pp. 5–17, 2009.

[4] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[5] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Quality and Safety*, vol. 28, no. 3, pp. 231–237, 2019.

[6] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10843 LNCS, pp. 593–607, 2018.

[7] F. Gong, M. Wang, H. Wang, S. Wang, and M. Liu, "SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation," *Big Data Research*, vol. 23, p. 100174, feb 2021.

[8] B. Hettige, W. Wang, Y. F. Li, S. Le, and W. Buntine, "MedGraph: Structural and temporal representation learning of electronic medical records," in *Frontiers in Artificial Intelligence and Applications*, vol. 325, pp. 1810–1817, 2020.

[9] A. J. Schaefer, M. D. Bailey, S. M. Shechter, and M. S. Roberts, "Modeling Medical Treatment Using Markov Decision Processes," *Operations Research and Health Care*, pp. 593–612, 2006.

[10] M. Gholinejad, A. J. Loeve, and J. Dankelman, "Surgical process modelling strategies: which method to choose for determining workflow?," *Minimally Invasive Therapy and Allied Technologies*, vol. 28, no. 2, pp. 91–104, 2019.

[11] Y. Zhang, P. Qi, and C. D. Manning, "Graph convolution over pruned dependency trees improves relation extraction," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 2205–2215, 2020.

[12] J. T. Senders, P. C. Staples, A. V. Karhade, M. M. Zaki, W. B. Gormley, M. L. Broekman, T. R. Smith, and O. Arnaout, "Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review," 2018.

[13] N. Prasad, L. F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," in *Uncertainty in Artificial Intelligence - Proceedings of the 33rd Conference, UAI 2017*, 2017.

[14] G. Lamé and M. Dixon-Woods, "Using clinical simulation to study how to improve quality and safety in healthcare," 2020.

[15] M. Srinivasan, J. C. Hwang, D. West, and P. M. Yellowlees, "Assessment of clinical skills using simulator technologies," 2006.

[16] C. Faudeux, A. Tran, A. Dupont, J. Desmontils, I. Montaudié, J. Bréaud, M. Braun, J. P. Fournier, E. Bérard, N. Berlengi, C. Schweitzer, H. Haas, H. Caci, A. Gatin, and L. Giovannini-Chami, "Development of Reliable and Validated Tools to Evaluate Technical Resuscitation Skills in a Pediatric Simulation Setting: Resuscitation and Emergency Simulation Checklist for Assessment in Pediatrics," *Journal of Pediatrics*, vol. 188, pp. 252–257.e6, 2017.

[17] S. Gershov, Y. Ringel, E. Dvir, T. Tsirilman, E. Ben Zvi, S. Braun, A. Raz, and S. Laufer, "Automatic Speech-Based Checklist for Medical Simulations," in *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, (Stroudsburg, PA, USA), pp. 30–34, Association for Computational Linguistics, 2021.

[18] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[19] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pp. 2463–2473, 2020.

[20] B. He, D. Zhou, J. Xiao, X. Jiang, Q. Liu, N. J. Yuan, and T. Xu, "BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models," pp. 2281–2290, 2020.

[21] F. Wu, T. Zhang, A. H. de Souza, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 11884–11894, 2019.

[22] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–14, 2017.

[23] L. Yao, C. Mao, and Y. Luo, "Graph Convolutional Networks for Text Detection," *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[24] R. Chandra, S. Goyal, and R. Gupta, "Evaluation of Deep Learning Models for Multi-Step Ahead Time Series Prediction," *IEEE Access*, vol. 9, no. May, pp. 83105–83123, 2021.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[26] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[27] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *7th Python in Science Conference (SciPy 2008)*, pp. 11–15, 2008.

[28] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

בהתבסס על תוצאות המחקר, המערכת שפיתחנו מזהה ומתעדת בצורה כמעט מדויקת את הבקשה להזרקת אדרנלין עם ציון ICC של 0.783. יתר על כן, המערכת אף הצליחה לחזות את תזמון הזרקת האדרנלין מהר יותר מהרופא האנושי, עם פערי זמנים של 00:01:28 לטובת המערכת שלנו. גם תחזית השימוש בדפיברילטור הניבה תוצאות מבטיחות. על ידי חיזוי של צעד אחד השגנו פער זמן של 00:00:50 דקות לטובת המערכת שלנו, בעוד חיזוי של שני צעדים השיגו פער זמן של 00:01:12 דקות. בנוסף, שתי שיטות החיזוי הנ"ל ניבאו בהצלחה את הצורך בדפיברילטור בסימולציה שבה הנבחן האנושי מעולם לא עשה זאת.

המשימות ברמת דיוק של $F_1 = 0.79$.

פיתוח מערכת ממוחשבת, מבוססת דיבור בלבד, לניהול סימולציות רפואיות תשפר את החוויה של מגוון רחב של פלטפורמות סימולציה. יתר על כן, מערכת ממוחשבת תוכל בהמשך הדרך לעזוב את עולם הסימולציה המבוקר ולהיות מיושמת בחדר הניתוחים ובמיון במטרה לסייע במצבי חירום רפואיים.

מערכת אוטומטית לחיזוי פרוצדורה רפואית על בסיס דיבור

במצבי חירום רפואיים בהם חייו של מטופל מוטלים על הכף, תהליך קבלת ההחלטות ושיקל דעתו של הרופא המרדים הוא החשוב ביותר. עם התמורות הטכנולוגיות בעולם הבינה המלאכותית, פותחו עם השנים מערכות תומכות החלטה קליניות (CDSS) ותפקידן הוא לסייע לצוות הרפואי בבית החולים. כיום, CDSS משמשים בעיקר לשיפור הידע של הרופא על ידי שילוב של הצעות המערכת CDSS עם מידע קליני המסופק למטפל. מערכות CDSS מסוגלות לבצע היקף עצום של פעולות, לרבות בקרת תרופות, מערכת אזעקה ותזכורות ועוד רבות אחרות. יתר על כן, מערכות אלו נדרשות לעבוד בזמן אמת בסביבות עבודה כאוטיות עמוסות הסחות דעת ורעשים וכן לתפקד במצבים של עמימות אבחנתית.

במחקר זה אנו מבקשים להעריך את הפוטנציאל של מערכת CDSS אוטונומית לחלוטין במצבי חירום רפואיים. בכדי שהמערכת תוכל לפעול בצורה אוטונומית, עליה להתאפיין ברמה מסוימת של מודעות והבנה על הנעשה סביב מיטת המטופל. למעשה, על המערכת להיות בעלת תבונה אשר לא תשפיע על עבודת הרופאים אלא במקרים בהם היא נדרשת.

בכדי שמערכת CDSS תהיה בעלת היכולת לפעול ולהשפיע במהלך מצבי חירום רפואיים, יש להשתית אותה על שני מרכיבי היסוד הבאים: ידע רפואי מוטמע ורציפות בקבלת החלטות. שילוב טכנולוגיות מעולם הבינה המלאכותית המכילות רכיבים מרכזיים אלו, יספקו תמיכה בהיקף ניכר של החלטות ובמיוחד במקרים של חוסר ודאות. בנוסף, מערכות אלו יוכלו לבחון ולהציע אלטרנטיבות לאופן הטיפול ולהעריך את ההשלכות של הטיפול המוצע על ידי הרופאים.

במסגרת המחקר שלנו יישמנו טכניקות מעולם הרשתות הגרפיות ומודלי השפה לצורך בניית תשתית למערכת CDSS אוטונומית לחלוטין, המבוססת על נתונים קליניים שנאספו מ - 28 סימולציות רפואיות עליהן פירטנו קודם לכן. הסיבה שבגינה אנו נעזרים בסימולציות רפואיות ולא במידע רפואי אמיתי טמונה בעובדה שאיסוף נתוני חירום רפואיים בבית חולים הוא משימה קשה בשל אופייה הבלתי צפוי והמורכב של סביבת העבודה. בנוסף, נדירותם של מצבים קליניים מסוימים הופכת את תהליך האיסוף לאקראי וארוך מידיי. סימולציה רפואית היא חלופה מתאימה לנתונים קליניים אמיתיים, מכיוון שהיא יכולה לספק את האפשרות לצפות ישירות באירועים בסביבה בטוחה ומבוקרת.

במסגרת מחקר זה יישמנו במערכת האוטונומית שני תפקידים חשובים של CDSS: בקרת תרופות והתערבות בזמן אמת. המערכת מבססת את החלטותיה בהתאם לניתוח שטף הדיבור של הצוות הרפואי, ומסוגלת לנבא תחזיות אודות אופן הטיפול אותו מבקשים הצוות הרפואי ליישם. בהתבסס על תרחישי הסימולציות, רצף המשימות הנדרשות משמשות לבניית מסלול של מודל העבודה האידיאלי. באמצעות אותה מערכת לייצור אוטומטי של רשימת משימות שבוצעו אנו יכולים לבנות עבור כל סימולציה את המסלול המתאים. מסלולי הפרוצדורות הרפואיות הם היסודות להטמעת ידע רפואי. לאחר מכן, עם שיטות מתקדמות של מודלי שפה ורשתות גרפים, המערכת המפותחת משתמשת במידע שנאסף, הפעולה הקודמת והנוכחית, כקלט לצורך חיזוי רצף הפעולות המלא בו ינקוט הרופא. על ידי צבירה מתמדת של מידע, המערכת מבצעת קבלת החלטות רציפה כדי להסיק את התחזית המדויקת ביותר בכל רגע נתון.

תקציר

מערכת אוטומטית לזיהוי פרוצדורה רפואית על בסיס דיבור

אימונים רפואיים הנעשים בסימולטורים רפואיים מספקים סביבה מבוקרת להכשרה והערכת כישורים קליניים. עם זאת, כפלטפורמת הערכה, היא דורשת נוכחות של בוחן מנוסה כדי לספק משוב על ביצועים, שנקבע בדרך כלל באמצעות רשימת משימות מפורטת. הצורך בבוחן אנושי הופך את תהליך ההערכה ליקר – הן מבחינת הזמן והן מבחינת העלויות. יתר על כן, שיטת הערכה זו אינה מספקת לרופאים מתמחים את הזדמנות לתרגול עצמאי. במבחן סימולציה אידאלי תהליך מילוי הרשימה צריך להיעשות על ידי מערכת אובייקטיבית בעלת מודעות קבועה ומוחלטת לסביבת הבחינה, המסוגלת לזהות ולפקח על הביצועים הקליניים של הנבחן. בכך המערכת מסייעת בהפחתת עלויות הערכת הביצועים ומאפשרת למתמחים להתאמן בתרחיש מורכב. בזכות הפיתוחים הטכנולוגיים שנעשו בשנים האחרונות בעולם מדעי המחשב, תהליך מילוי הרשימה יכול להתבצע על ידי מכונה - מערכת אובייקטיבית מודעת לחלוטין המסוגלת לזהות ולפקח על ביצועי הנבחנים. במסגרת מחקר זה אנו טוענים את הנחת היסוד הבאה - במקרים רבים התקשורת בין הצוותים הרפואיים עשויה לייצג את הפעולה הרפואית עצמה. על ידי ניתוח שטף דיבור המשתתפים, אנו יכולים לזהות ולסמן המשימות המתאימות ברשימה. לשם כך פיתחנו מערכת אוטומטית מבוססת דיבור בלבד המזהה את פעולות המתמחים ומבצעת בצורה אובייקטיבית את אימות המשימות.

המערכת נבדקה באמצעות שני תרחישים קליניים שונים להערכת כישורי מתמחי הרדמה בכירים. שני התרחישים פותחו על ידי רופא מרדים מנוסה ומומחה לסימולציה רפואית. התרחישים התבססו על תרחישים שנכתבו בעבר על ידי הרופא המרדים (א.ר) ושימשו לבחינת ההסמכה של מועצת ההרדמה הישראלית. התרחיש הראשון כלל טיפול בחולה עם תגובה אנפילקטית חמורה והתרחיש השני כלל מטופל לאחר ניתוח הסובל מברדיקרדיה קשה.

20 מתמחים בכירים להרדמה, 13 גברים ו - 7 נשים, השתתפו במחקר. בנוסף, שני חברי צוות המחקר מילאו את תפקידי האחות ומתמחה ברפואה המסייעים לנבחן. במהלך הסימולציה, רופא מרדים מנוסה העריך את ביצועי הנבחן באמצעות רשימת הבידוק המתאימה לתרחיש. המטופל בכל אחת מהסימולציות הייתה בובת גוף מלאה המיועדת לתרגול מסוג Kelly MegaCode של חברת 'Laerdal'. כמו כן, בכדי לגרום לתרחיש להיראות אמיתי ככל האפשר פיתחנו מערכת לאיסוף נתונים בתוך מחלקת ההתאוששת של בית החולים רמב"ם. וידאו ושמע הוקלטו באמצעות תוכנת הקלטת וידאו דיגיטלית (NorPix StreamPix Inc.). נתוני הווידאו שהוקלטו שימשו את הצופה האנושי למילוי ידני של רשימת המשימות. לצורך הקלטות שמע, כלל משתתפי הסימולציה לבשו משדר מיקרופון אלחוטי כאשר כל ערוץ שמע נשמר בנפרד. המחקר אושר על ידי ועדת המרכז הרפואי רמב"ם. התהליך האוטומטי של ייצור רשימת המשימות שבוצעו כלל מספר שלבים. תחילה בוצע תמלול אוטומטי של כל אחד ממשתתפי הסימולציה (3 משתתפים), ולאחר מכן בכל אחד מן המשפטים המתומללים בוצע חיפוש וזיהוי של מילות מפתח. באמצעות מילות מפתח אלה יושם תהליך התאמה בין המשימות אותן צריך היה הנבחן לבצע לבין המשפטים שנאמרו. תוצאת האלגוריתם הייתה רשימת משימות מלאה שבה לצד כל משימה שבוצעה הוצמד משפט מתומלל תואם וחותמת זמן. בהתבסס על התוצאות שהתקבלו, המערכת מסוגלת לזהות את רוב

המחקר בוצע בהנחייתו של ד"ר שלומי לויפר בפקולטה לתעשייה וניהול.

חלק מן התוצאות בחיבור זה פורסמו כמאמרים מאת המחבר ושותפיו למחקר בכנסים ובכתבי-עת במהלך תקופת מחקר המאסטר של המחבר, אשר גרסאותיהם העדכניות ביותר הינן:

Sapir Gershov, Yaniv Ringel, Erez Dvir, Tzvia Tsirilman, Elad Ben Zvi, Sandra Braun, Aeyal Raz, and Shlomi Laufer. Automatic Speech-Based Checklist for Medical Simulations. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 30–34, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

תודות

בראש ובראשונה, ברצוני להביע את תודתי הכנה למנחה שלי ד"ר שלומי לויפר על התמיכה המתמשכת במחקר שלי ב- MS.c, על סבלנותו, המוטיבציה, ההתלהבות והידע העצום שלו. ההנחייה שלו עזרה לי בכל זמן המחקר והכתיבה של עבודת תזה זו. לא יכולתי לדמיין יועץ ומנטור טוב יותר. מלבד המנחה שלי, אני רוצה להודות לד"ר אייל רז על העידוד, ההערות, התובנות והשאלות הקשות.
תודתי הכנה מגיעה גם לד"ר מורן ארצי שהאירה לי את הדרך במחקר הראשון שלי ולימדה אותי כיצד לשתות קפה טורקי.
אחרון חביב, ברצוני להודות להורי ולאחיי, על כך שתמכו בי לאורך כל חיי ותמיד עוררו בי השראה להגיע לשיאים חדשים.

הכרת תודה מסורה לתוכנית הבין-יחידתית לרובוטיקה ומערכות אוטונומיות על מימון מחקר זה.

מערכת אוטונומית לסימולציות רפואיות

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
מגיסטר למדעים במערכות אוטונומיות ורובוטיקה

ספיר גרשוב

הוגש לסנט הטכניון — מכון טכנולוגי לישראל
חשוון תשפ"ב חיפה אוקטובר 2021

מערכת אוטונומית לסימולציות רפואיות

ספיר גרשוב