Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Xuanyang Lin April 8, 2023

Multi-task Multi-sensor Framework for Assessing Health Effects of Heat Exposure with Medical Sensors

by

Xuanyang Lin

Dr. Li Xiong
Adviser

Computer Science

Dr. Li Xiong

Adviser

Dr. Jinho Choi

Committee Member

Dr. Vicki Hertzberg

Committee Member

2023

Multi-task Multi-sensor Framework for Assessing Health Effects of Heat Exposure with Medical Sensors

By

Xuanyang Lin

Dr. Li Xiong

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science

2023

Abstract

Multi-task Multi-sensor Framework for Assessing Health Effects of Heat Exposure with Medical Sensors

By Xuanyang Lin

Medical sensors help to monitor and collect data that make early treatment decisions possible. Using the well-collected data from these sensors, recent studies have focused on the early predictions for some life-threatening conditions, such as dehydration and kidney injuries, in order for early intervention and prevention. In previous studies, the use of multi-sensor fusion transformer models has been proved to be very effective in fusing multi-modal data and detecting signals for various classification tasks. Since many of these tasks are closely related, it is reasonable to incorporate them into a multi-task learning (MTL) framework, so that task-specific information can be shared across different tasks, and that the training and inference cost could be brought down significantly. In this study, we propose a multi-task multi-modal transformer framework that handles several medical predictions. We combine task-specific losses by a dynamic weighting strategy that balances individual losses. To better tackle the label noise problems of our dataset, we also incorporate a teacher-free regularization method (Tf-KD) into our framework. We evaluate the method on the classifications of acute kidney injury (AKI) and dehydration (USG) on Girasoles sensor dataset. We find that MTL benefits both tasks, while Tf-KD only helps the prediction of AKI, suggesting that further research needs to be done.

Multi-task Multi-sensor Framework for Assessing Health Effects of Heat Exposure with Medical Sensors

By

Xuanyang Lin

Dr. Li Xiong

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science

2023

Acknowledgements

I would like to express my gratitude to Dr. Xiong, who proposed the multi-task multi-sensor project and who has supported me throughout my entire research experience. My work is also impossible without Dr. Rongmei Lin, whose previous work set the foundation of my method and who mentored me in this project. I also want to say thanks to Dr. Jinho Choi and Dr. Vicki Hertzberg, my honor thesis committee members. Their evaluation of my work is very crucial, and their feedback helps me improve this thesis and suggests how I can improve my work in the future.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The health effects of environmental heat exposure are the most significant cause of weather-related mortality in the U.S [15]. In the field of agriculture, over 750,000 workers are exposed to hot weather on a daily basis, leading to a very high mortality due to heat-related illnesses (HRI). However, HRI is entirely preventable and is also treatable provided it is detected in a timely manner [13]. Recently, the applications of wearable monitors and electronic health records (EHR) enable us to collect large quantities of high-frequency data with a variety of signals. In particular, multi-sensor data collection in farmers gathers time-series data of various modalities. This advancement of data availability triggered an effort to analyze the signal patterns in order to predict possible diseases or conditions that are life-threatening, such as acute kidney injury (AKI) and dehydration (USG) [16].

Each sensor collects features that are different from other sensors, so the features are multi-modal. In the study of processing multi-modal sensor information, early fusion and late fusion are most commonly used methods [8]. In early fusion, signals from different modalities are pre-processed and concatenated in the early phrase. Fea-

tures are extracted from such combined signals and feed into the downstream task like classification. In late fusion, raw signals from each sensors are featurized separately and then fused for downstream task [13]. Besides early and late fusions, intermediate fusion is a data fusion method that is gaining traction recently. It start by extracting features for individual modalities, and then combine modalities in cross-modality layers.

1.1 Challenges

Despite previous efforts to model the complexity of multi-modal sensor data, there are several challenges in modeling the medical sensor data and construct accurate models.

Firstly, capturing both the intra- and inter- modality dependencies is still a challenge. Early fusion fails to process intra- and inter- modality relationships differently because all the features are concatenated in a very early phase. Late fusion cannot capture inter-modal information because extracted features from each modality are not further processed in some common neural network architecture.

Secondly, in medical sensor data, prediction tasks can be highly related to each other. In specific, since HRIs are triggered by heat, the occurrences of them are inherently related to core body temperature. Such a task similarity presents an opportunity to improve the predictive power by doing multi-task learning (MTL) [5], a modeling technique in which multiple tasks are trained in a single model architecture to share task-specific domain knowledge. For example, the classifications of dehydration and body temperature are highly correlated to each other. To incorporate MTL into the predicative model, we need to find a great parameter-sharing strat-

egy by properly designing the MTL model structure. Besides, the possibility of task imbalances necessitates a method to combine task-specific losses.

Lastly, we find that certain prediction tasks are inflicted by label noise. Because of the difficulty in data collection, some tasks have a very low sampling rate, while others are sampled frequently. For instance, the collection of AKI labels needs blood test, so they are usually collected only once a day, while the core body temperature (Temp) is easy to measure and is measured per 30 sec. To fill in the gap of sampling rate differences, we use closest measurements to impute on the missing values. This imputation procedure imply some labeling error, so the task is inflicted by label noise, which can cause the model to overfit.

1.2 Contributions

To tackle the difficulty in handling multi-modality data, a recent study [13] proposed a **multi-sensor fusion framework** consist of modality-specific gated recurrent units (GRUs) and multi-sensor transformer. This model is much more flexible and is capable of capturing both intra- and inter- modality correlations. It has been shown to work well for classification tasks. In our study, we use this method as the basis for modeling complex multi-modal sensor data.

Based on the multi-sensor fusion [13], we design a MTL framework that incorporates two or more classification tasks, in which we select similar tasks to be trained in one model structure. As an example, we have core body temperature as an auxiliary task to help the classification of AKI. We employ a hard-parameter-sharing strategy in which the multi-sensor fusion architecture serves as the backbone. And we use a dynamic weighting of task-specific losses to balance the training of tasks.

We also add a regularization method for certain noisy tasks to combat overfitting. After comparing different regularization methods and experimenting on some of them,

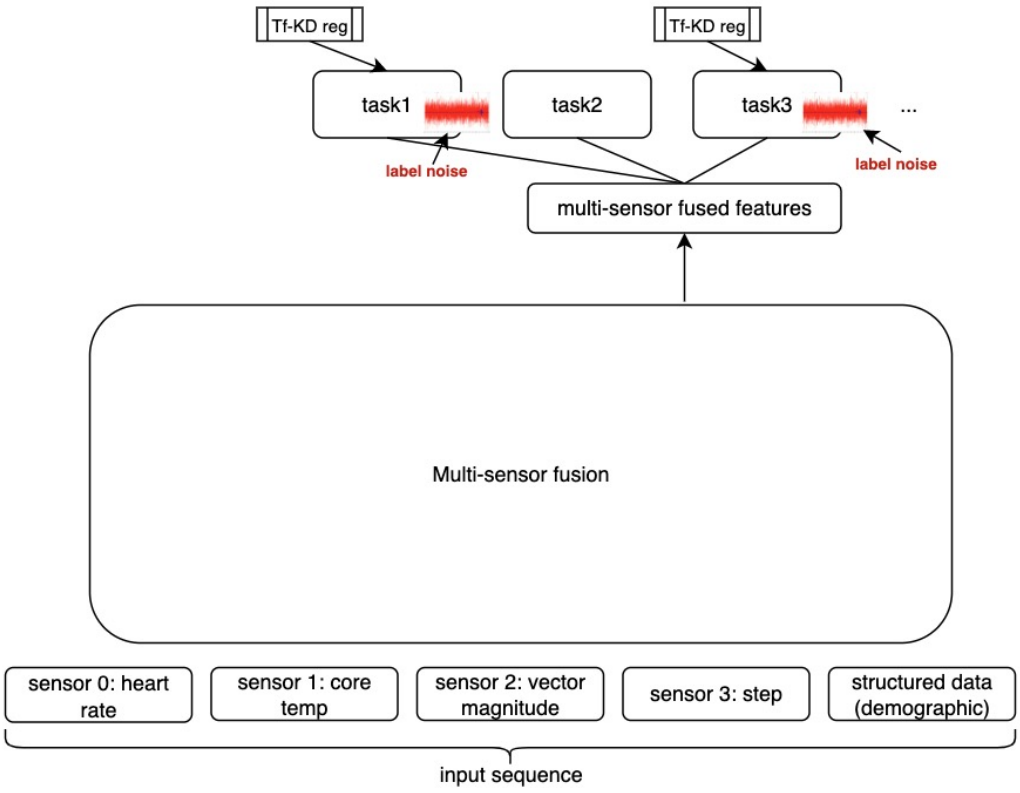

Figure 1.1: Multi-task Multi-Sensor Fusion Framework.

we find teacher-free knowledge distillation (Tf-KD) [20] as an optimal regularization method. For noisy tasks, we add this regularization term to their original cross entropy losses.

Combining multi-sensor fusion, MTL, and Tf-KD, we propose a multi-task multi-sensor framework as shown in Figure 1.1. We conduct extensive experiments on the Girasoles sensor dataset [16] that collects thousands of hours of sensor signals of agricultural workers in Florida. In specific, we use heart rate, core body temperature, mobility as predictors to classify the occurrences of acute kidney injury (AKI) and dehydration (USG), and core temperature (if the core body temperature will exceed 38.0∘C in the next 10 minutes or not) . We evaluate the effectiveness of three components individually and the effectiveness of the framework as a whole.

Chapter 2

Background

2.1 Multi-modal fusion

In early fusion, different modalities, either raw or pre-processed, are combined into single representation by concatenation, which is followed by feature extraction. For example, early fusion was used to fuse audio and image data to produce common representations [3]. But,early fusion does not handle the difference between intra- and inter- modal dependencies, because modalities are combined at the first stage, before any feature extraction. Besides, this method requires innovations in sensor synchronization, buffering, denoising and data normalization.

Late fusion, on the other hand, extracts features by modality-specific feature extractors or classifiers. Outputs of each modality are combined and assembled at the very last step to produce the final prediction. As an example, deep late fusion neural network has been applied on image and audio representations for classifications [4]. This fusion method, however, cannot effectively extract inter-modality relationships, since the fusion of modalities happens all feature extractions.

Recently, intermediate fusion, in which modalities are fused midway, is gaining traction. Its use of modality-specific networks and cross-modal fusion network is more

likely to capture more comprehensive information in multi-modal data. In the field of visual question answering (VQA), for example, a study processes and fuses various modalities (question words, object images, etc.) by first extracting intra-modality features with appropriate models and then combining modalities in a few attention layers [10]. In the field of disease prediction with medical data, MultiFusionNet [19] combines two sources of knowledge, extracted features and raw data, by two separate deep neural networks followed by cross-domain layers.

Intermediate fusion is generally a well-developed fusion method that could potentially be a good fit for medical multi-sensor analysis, but the design of cross-modal fusion network is very important. One of a commonly-used method, gated fusion, uses gated attention to fuse multi-modal features, but the attention is only allocated on modality-level, not on each individual feature [13]. In medical sensor, a study By Rongmei Lin [13] uses intermediate fusion method with a two-step feature extraction process to learn dependency within modalities and fuse features across modalities. It applies gated recurrent units (GRUs) [6] for each modality, and a cross-modality transformer to learn the complex inter-modality interactions [13]. This method can be considered as an integration of early and late fusions, with both modality-specific feature extraction and cross-modality layers, and it is more flexible and superior to early and late fusions. It is also superior to gated fusion by more a flexible attention mechanism in cross-modality transformer. In our study, we use this work as the basis of our method.

2.2 Multi-task learning

Multi-task learning (MTL) has been proved to be a great method to improve multiple tasks simultaneously and save computation resources by using less parameters [18]. When we have two or more (predictive or generative) tasks that are similar to each

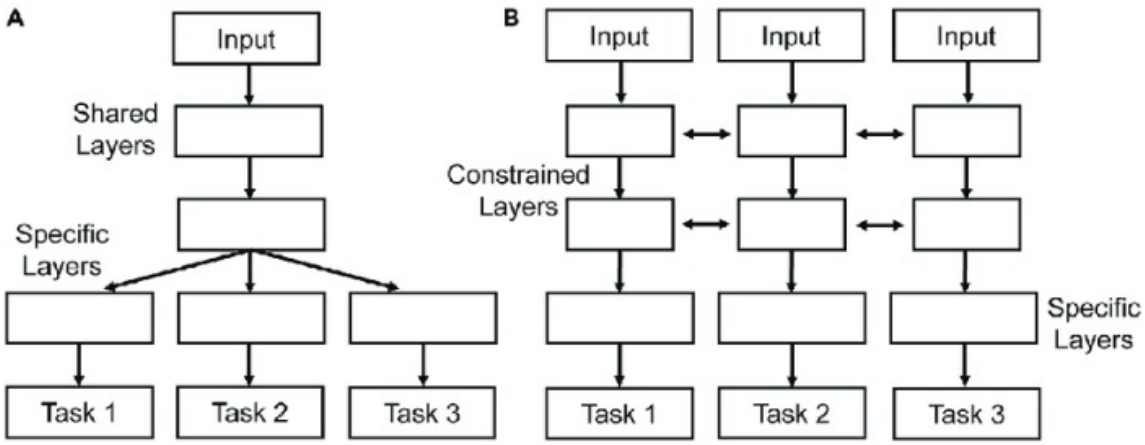

Figure 2.1: Hard parameter sharing (left) and soft parameter sharing (right) [7].

other, MTL shares task-specific knowledge so that each task can access not only its own information, but also information from other tasks. Besides, MTL can be considered as a form of regularization method [18]. While different task has different noise, a MTL architecture filters out task-specific noise and extracts more generalizable features.

MTL methods can be categorized into two classes: hard parameter sharing and soft parameter sharing [18] (Figure 2.1). Hard parameter sharing MTL is the most common method, in which a few common layers are shared by all tasks, followed by some task-specific layers. The model learns a common representation for various tasks and produce outputs for each task by task specific heads for classification, regression, or generation. Soft parameter sharing MTL does not have any layers being shared by multiple tasks. Instead, each task has its own set of layers, and the parameters are regularized to encourage the parameters of different tasks to be similar. As MTL has great success on many problems, this study applies a carefully-designed hard-parameter-sharing [**?**] MTL learning algorithm on medical sensor dataset with the multi-sensor fusion method, with a dynamic weighting strategy of losses [14].

2.3 Knowledge distillation as regularization

Model overfitting is a common issue posing serious challenges to an effective machine learning algorithm. To combat model overfitting, common methods include data augmentation, dropout, weight decay, and L1 and L2 regularization. However, a recent study [17] tests the effects of many of these methods and found that they have limited effect.

Knowledge distillation (KD) [9] is a very effective method to transfer the information from a large strong network (teacher) to a smaller network (student model), enabling the student to achieve similar predictive power as the teacher without bearing a large number of parameters. During the training of the student, the KL-divergence term between the model probability of the student and that of the pre-trained teacher is added to the original loss, encouraging the student output imitate the teacher output [9]. In order for the KD to work, this study on KD [9] implies that the teacher model has to be more knowledgeable about the task. The loss function for a classification can be expressed as

$$L = (1 - \alpha)H(q, p) + \alpha D_{KL}(p^t, p)$$

in which H is the original cross entropy, D_{KL} is the KL-divergence, q, p, p^t are ground truth, student output, and teacher output, respectively [9].

While a typical KD assumes that the teacher needs to be stronger than the student, a recent study [20] found that a weakly-trained teacher still significantly improves the student. The study proposes that KD has a regularization effect by establishing a link between KD and label smoothing (LS). For LS, Yuan et al. (2021) found that the loss function can be split into the loss function using labels without smoothing (one-hot

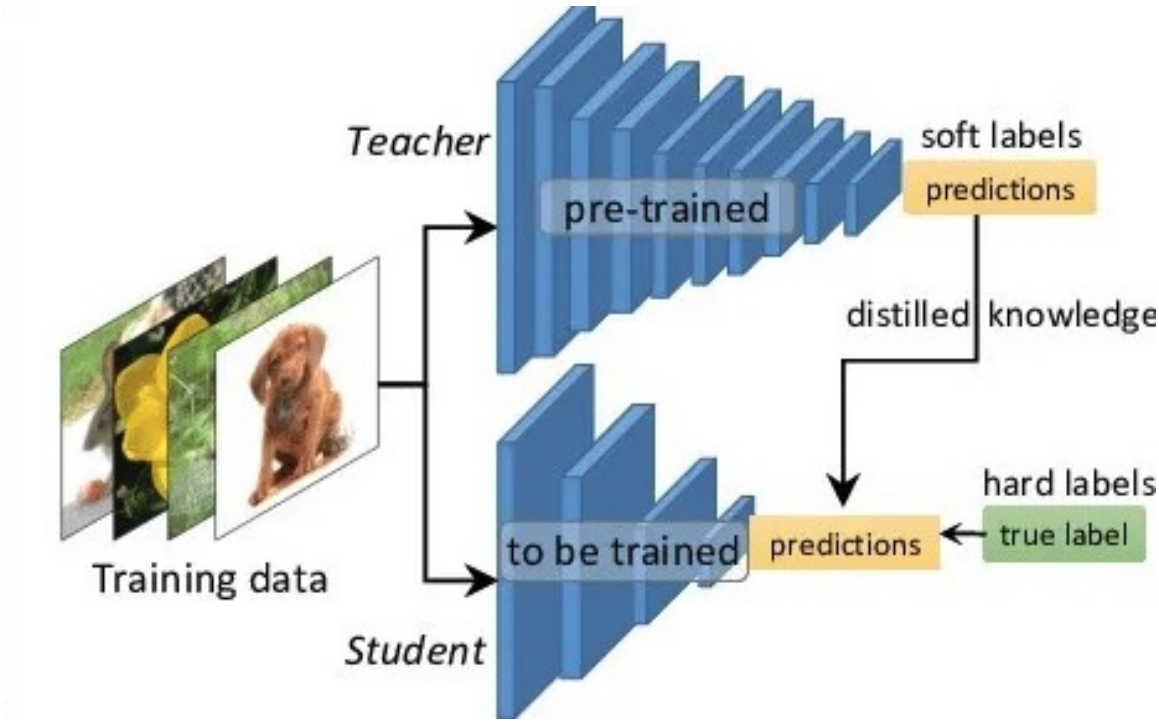

Figure 2.2: Knowledge Distillation [2].

labels) and a KL-divergence between the model output and an uniform distribution:

$$L = (1 - \alpha)H(q, p) + \alpha D_{KL}(u, p)$$

So, LS can be viewed as an ad-hoc KD with a random classifier as teacher. For KD, the study found that KD is a special case of LS where the label smoothing distribution is a learned distribution from the teacher. Since LS is effectively a regularization method by preventing the overconfidence of model probability, Yuan et al. (2021) establishes KD as a regularization, with potentially stronger effects due to the fact that the smoothing distribution is a learned distribution [20].

As KD has been proved to be a regularization method, the study in [20] proposed an innovative KD method (Tf-KD) that operates without a teacher. Without an extra model to regularize, we can still apply KD regularization by building a "virtual teacher" whose output distribution is manually designed [20]. Previous study has

been focusing on Tf-KD regularization on single-task learning models [20], so we would like to adopt Tf-KD in our MTL framework to see if Tf-KD works well in MTL context.

Chapter 3

Proposed method

3.1 Multi-sensor fusion

We use the multi-sensor fusion framework [13] as the basis of our method, as shown in Figure 3.1. In this framework, features of different modalities are first processed in modality-specific networks that capture dependencies within each modality, while demographic information is processed by a feed-forward network. In multi-sensor fusion [16], the study chooses GRU over other recurrent neural networks (RNN), since GRU's reset and update gates help to capture temporal dependency more effectively [6], while its training is much simpler than long short-term memory (LSTM). Then, the extracted features are passed into a cross-modality transformer that learns inter-modality relationships. By learning both intra-modality and inter-modality signals, this multi-modal framework achieves high performance [13].

In this study, we use the multi-sensor fusion, with its modality-specific GRUs and multi-sensor transformer, as the model backbone that is shared by all tasks.

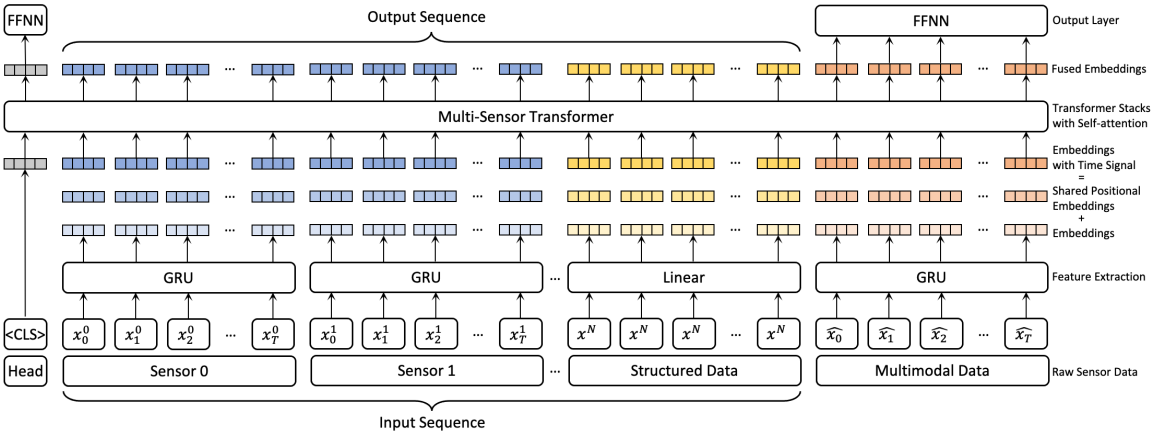

Figure 3.1: Multi-Sensor Fusion Framework [13].

3.2 Multi-task learning

In our multi-task multi-sensor framework, we apply a hard-parameter-sharing strategy for learning multiple tasks simultaneously [5]. In particular, the modality-specific GRUs and the multi-sensor transformer serve as the backbone of our framework shared by all tasks. On top of the backbone, the extracted features would then be used as inputs for independent task-specific heads consisting of fully-connected layers that output model probabilities.

When we add up task-specific losses into one, it is imperative to balance the losses so that every task can make progress in training. We apply a dynamic weighting strategy in which the loss weights are negatively correlated with the progress of loss reduction in the last training iteration, so that tasks with less progress will be of higher focus in the current iteration [14]. Let λ_i be the weight for task i for current iteration, L_i^{t-1} and L_i^{t-2} the values of loss at previous two steps, respectively. The ratio between L_i^{t-1} and L_i^{t-2} represents the progress on this specific task at last step:

$$r_i = \frac{L_i^{t-2}}{L_i^{t-1}}$$

We use this ratio to weight task-specific losses ($\lambda_i = r_i$). The smaller this ratio,

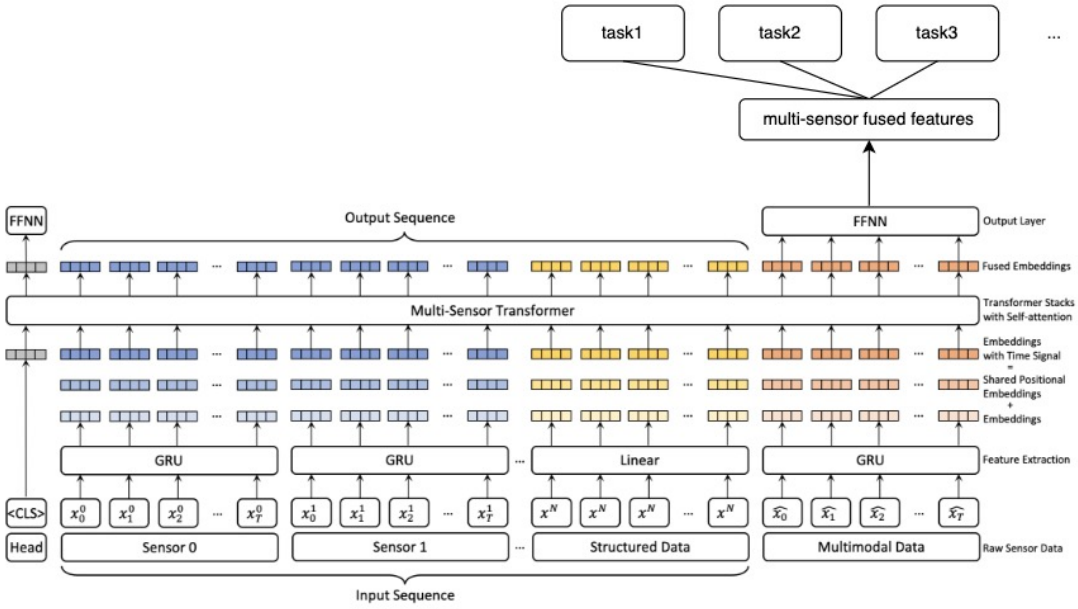

Figure 3.2: Multi-Sensor Fusion Framework with MTL [13].

the less progress that the task made on last step, and the more it needs to be focused on on current step. So, the weighted loss for our MTL framework with n tasks is as follow:

$$L = \sum_{i=0}^{n} \lambda_i L_i$$

3.3 Teacher-free knowledge distillation

As described previously, knowledge distillation (KD) is a great regularization method similar, but potentially superior, to label smoothing (LS). To apply KD regularization (KDR) method without an additional teacher model, one can always design a "virtual teacher" by manually designing a distribution, a method called teacher-free knowledge distillation (Tf-KD) [20]. As we found that some of the tasks suffer from overfitting, we believe that Tf-KD can improve the overall performance on evaluation.

We apply Tf-KD on noisy tasks as a regularization term. Following the design of the study [20], we build a simple virtual teacher by manually designing its output

distribution for classes as the following [20]:

$$p^{teacher}(k) = \begin{cases} a & \text{if} \quad k = c \\ (1-a)/(K-1) & \text{if} \quad k \neq c \end{cases}$$

where K is the total number of classes (K=2), c is the correct label and a is the probability for the correct class. As suggested by the study [20], we set a=0.99 (this is a hyperparameter that should be tuned in our future work), so that the probability for the correct class is much larger than that of the incorrect class, which means that the virtual teacher will always make right predictions. We then add a KL-divergence term to the original cross entropy loss function [20]:

$$L = (1-\alpha)H(q,p) + \alpha D_{KL}(p^{teacher}, p)$$

where p is the model output, $p^{teacher}$ is the virtual teacher output, and q is the ground truth. α is the hyperparameter that controls the weighting between cross entropy and KL-divergence terms. We set it to be 0.5, as we find by experiments that it generally gives good results.

Chapter 4

Experiments

We use the multi-sensor fusion model [13] as our baseline, and evaluate it on the classifications of AKI and USG. In this study, we conduct three groups of experiments to evaluate the effectiveness of MTL, Tf-KD, and a combination of the two methods. In our experiments, we first conduct MTL evaluations to test the effectiveness of MTL with two questions in mind: 1) does MTL of AKI or USG with Temp improves AKI or USG; 2) is there any differences among MTL strategies (e.g. are AKI & Temp MTL and AKI & USG & Temp different). Then, we apply the Tf-KD regularization on AKI and USG to test its effects on the tasks. Lastly, we incorporate both MTL and Tf-KD into the multi-task multi-sensor framework to see if the combined model further improves the performance. We compare the results against the multi-sensor fusion model as baseline in the Girasoles Sensor dataset [16].

4.1 Experimental setup

Model Architecture. We use the multi-sensor fusion [13] as our baseline, and the evaluation of it on evaluation metrics (marco and micro AUROC, accuracy) is presented in Table 4.1. For the multi-task multi-sensor framework, the structure of the backbone (GRUs and multi-sensor transformer) is in align with the work of

multi-sensor fusion [13], with 4 stacks of transformer layers, 4 attention heads, and an embedding space dimension of 256.

Parameter setting. We use batch size of 64 during training. The model is trained using the Adam optimizer with initial learning rate 0.0001 and early stopping. For Tf-KD, we set the probability for correct class, a, to be 0.99, in align with the experiment setting of the Tf-KD study [20], and we set the weighting between KL-divergence and the cross entropy to be uniform, which we found to be most effective by a experiments. Further hyperparameter tuning needs to be done, especially on the correct class probability a.

Evaluation metrics. In this study, we consider three metrics: accuracy, micro and macro AUROC. While accuracy is simply defined as the proportion of correct classification, the definitions of micro and macro AUROC are more complicated and require more description. The micro AUROC computes its true positive rate (TPR) and false positive rate (FPR) by micro averaging aver classes as follows [1]:

$$TPR = \frac{\sum_c TP_c}{\sum_c (TP_c + FN_c)}$$

$$FPR = \frac{\sum_c FP_c}{\sum_c (FP_c + TN_c)}$$

While the macro AUROC requires computing TPR and FPR independently for each class and then taking the average over them [1].

$$TPR = \sum_c TPR$$

$$FPR = \sum_c FPR$$

While accuracy gives a rough measure of the effectiveness of the classification

model, AUROC evaluates the model under many different decision thresholds. The macro AUROC uses a macro averaging to compute the metric by calculating for individual classes and taking the average, so that the minority class has a larger weighting on the final score. The macro AUROC is an ideal metric for datasets with imbalanced classes, and as the Girasole sensor dataset is imbalanced, macro AUROC is more reflective of model performance.

Table 4.1: Performance of Multi-Sensor Fusion Model (Baseline)

task	macro AUROC	micro AUROC	accuracy
AKI	0.7227	0.9255	0.8626
USG	0.7033	0.9326	0.8745
Temp	0.9823	0.9937	0.9644

4.2 Girasoles sensor dataset

The Girasoles Sensor dataset captures 4,000 hours of core body temperature (per 30 sec, degree Celsius, two decimals precision), heart rate (per 30 sec, beats per minute), motion activity (per 60 sec, including vector magnitude and number of steps per minute), and occurrence of acute kidney injury (per day) and dehydration (per 12 hr) among 254 agricultural workers [16]. A pseudo sample of the dataset is shown in Table 4.2, where each row represents one record. The AKI is measured by serum creatinine in blood sample and urine sample using KDIGO guideline [16]. The occurrence of dehydration is measured by the urine sample if urine specific gravity (USG) is greater than 1.030 [16].

We consider heart rate, core temperature, vector magnitude and steps as input modalities. We also incorporate demographic information as a separate modality with structured data [13].

Train/test split. We split workers for train/test purpose, where approximately 20 percent of them are placed in the test group, and then randomly sampled consec-

Table 4.2: Pseudo sample of Girasoles sensor dataset.

time	hr	core temp	vector magnitude	step
7:00:00	101.9	36.98	1147.01	18
7:00:30	113.8	37.01	1147.01	18
7:01:00	117.4	38.19	1988.86	10
7:01:30	109.4	38.01	1988.86	10
7:02:00	117.0	37.99	1801.73	9
7:02:30	133.9	37.18	1801.73	9

utive windows of 50 minutes duration and generated 21,361 training data and 5,371 testing data [13].

Tasks. Our model aims to perform well on AKI and USG classifications, as they bring serious health consequences and are sometimes life-threatening. Between these tasks, AKI is generally more concerning, as it causes long-term damages to kidney functions. In addition, in order to announce the high temperature warning in a timely manner, we design an extra task to predict whether the individual's core body temperature will exceed $38.0°C$ in the next 10 minutes [13]. The real-time classification is applied in the last task. As previously mentioned, some of the tasks are very noisy. Due to the complexity of measuring AKI and USG occurrences that involves blood tests, the frequency of AKI and USG measurements are once a day and twice a day, respectively [16]. because of the difference in sampling frequencies, the vast majority of core body temperature data points are not matched by AKI or USG measurements, so we have to fill up the gap by using the measurements closest to them instead. Besides, labels for both tasks are imbalanced, with the AKI and USG ratios being 9.7 percent and 9.8 percent, respectively.

4.3 Multi-task learning

We first evaluate the effectiveness of multi-task learning AKI or USG with the auxiliary task (core temperature), as well as the differences among various task combi-

nations. We perform experiments that involve 3 MTL task combinations: AKI & Temp, USG & Temp, and AKI & USG & Temp. We compare all results to that of the multi-sensor fusion model, which serves as our baseline. Tables 4.3 and 4.4 below shows the accuracy and micro and macro AUROC on different task combinations, showing that all MTL combinations are effective, but two-task MTL (AKI & Temp, USG & Temp) performs better than 3-task MTL.

The performance on various MTL task combinations differ, showing Temp as the most helpful task and AKI and USG less helpful and even detrimental to each other. One possible explanation of 2-task over 3-task models is that both AKI and USG are directly related to high body temperature and excessive heating, whereas AKI and USG are less connected to each other. Besides, whereas AKI and USG labels involves imputation of missing data, all Temp labels are from real measurements, containing more valid information that may transfer to other tasks.

Besides main experiments, regarding the architecture of MTL models, we do a comparative study to investigate the parameter sharing strategy and see if the amount of shared model parameters may affect the model performance. The way in which our prediction tasks share parameters matters, because it determines if tasks are sharing low-level or high-level features. Apart from the current parameter sharing scheme where all GRUs and the multi-sensor transformer are shared, we also experiment on separating out the decoder as task-specific. However, by experiments, we did not observe any significant differences performance-wise. Because separating out decoders as task-specific increases the number of model parameters, we determine that the original model is superior, so we do not adopt this method.

Table 4.3: Performance of **AKI** in Multi-Task Learning

task	macro AUROC	micro AUROC	accuracy
AKI (baseline)	0.7227	0.9255	0.8626
AKI and Temp	0.8962	0.9885	0.9531
AKI, USG and Temp	0.8339	0.9700	0.9219

Table 4.4: Performance of **USG** in Multi-Task Learning

task	macro AUROC	micro AUROC	accuracy
USG (baseline)	0.7033	0.9326	0.8745
USG and Temp	0.9683	0.9856	0.9219
AKI, USG and Temp	0.8871	0.9922	0.9688

4.4 Teacher-free knowledge distillation (Tf-KD)

To test the effectiveness of Tf-KD, we evaluate single-task models with Tf-KD regularization. Tables 4.5, 4.6 and 4.7 shows the accuracy, micro and macro AUROC scores for AKI and USG, under Tf-KD setting. The results on AKI, the task of highest importance, show an almost 2 percent improvement in AUROC, although the accuracy is one percent lower. As AUROC is a more robust metric that evaluates the performance of the model under different threshold, we believe that Tf-KD is still taking a step forward.

However, results on USG do not show any progress. This may be explained by the fact that USG measurements are taken twice as frequent as measurements on AKI, so the USG task may be much less noisy. As Tf-KD is a method that targets label noise, we suspect that it only works well on high-noisy tasks. We also did Tf-KD experiments on Temp task, and the metrics do not show any improvement either.

Table 4.5: Performance of AKI with Knowledge Distillation

task	macro AUROC	micro AUROC	accuracy
AKI (baseline)	0.7227	0.9255	0.8626
AKI (Tf-KD)	0.7419	0.9275	0.8523

Table 4.6: Performance of USG with Knowledge Distillation

task	macro AUROC	micro AUROC	accuracy
USG (baseline)	0.7033	0.9326	0.8745
USG (Tf-KD)	0.6902	0.9316	0.8767

Table 4.7: Performance of Temp with Knowledge Distillation

task	macro AUROC	micro AUROC	accuracy
Temp (baseline)	0.9823	0.9937	0.9644
Temp (Tf-KD)	0.9815	0.9937	0.9650

4.5 Multi-task learning and Tf-KD regularization

Lastly, we experiment on a combination of MTL and Tf-KD, MTL with Tf-KD as regularization, to see if it outperforms MTL-only and KD-only settings. Since we find out in Tables 4.3 and 4.4 that Temp is helpful for AKI and USG and that AKI and USG are not mutually beneficial, we only test for AKI Temp & and USG & Temp with KD regularization. We test AKI-Temp and USG-Temp MTL, with Tf-KD applied on AKI and USG tasks, respectively. Tables 4.8 and 4.9 shows the accuracy, micro and macro AUROC for AKI-Temp (Tf-KD) and USG-Temp (Tf-KD), with results on MTL-only and Tf-KD-only models on the side for comparison. Again, the metric scores on AKI-Temp MTL show promising results, outperforming both MTL-only and KD-only models. Results on USG show better performance than the baseline, but the MTL-Tf-KD model does not outperform MTL-only one, implying ineffectiveness of Tf-KD regularization in dealing with USG.

Table 4.8: Performance of AKI in MTL and Tf-KD regularization

task	macro AUROC	micro AUROC	accuracy
AKI (baseline)	0.7227	0.9255	0.8626
AKI and Temp	0.8962	0.9885	0.9531
AKI (Tf-KD)	0.7419	0.9275	0.8523
AKI and Temp (Tf-KD)	0.9677	0.9971	0.9688

Table 4.9: Performance of USG in MTL and Tf-KD regularization

task	macro AUROC	micro AUROC	accuracy
USG (baseline)	0.7033	0.9326	0.8745
USG and Temp	0.9683	0.9856	0.9219
USG (Tf-KD)	0.6902	0.9316	0.8767
USG and Temp (Tf-KD)	0.85	0.9248	0.9375

Chapter 5

Conclusion

As a great attempt to assist the medical study with data-driven solutions, medical sensor models are an important branch of artificial intelligence study. 1) In this study, we handle the multi-modality of the data by applying multi-sensor fusion strategies [13]. 2) Based on the multi-sensor framework, we build effective multi-tasking strategies to take advantage of the commonality of various tasks. 3) The Tf-KD regularization [20] is applied on noisy tasks, and is proved to be effective in AKI, the most important task that we would like to secure. We combine all three components above into an unified multi-task multi-sensor fusion framework. While the classification of AKI is improved by both MTL and Tf-KD, the prediction on USG is not effective on our KD regularization method. Further research on USG is necessary.

Chapter 6

Future study

Although our method is generally effective in AKI, and partially effective in USG, we believe that the methodology can be further improved in various ways. This suggests some future study.

Firstly, we will explore more dynamic weighting strategy methods for weighting losses. We have experimented on several dynamic weighting strategies, including methods based on training loss reduction, absolute values of training losses or validation losses [14]. However, the currently best dynamic weighting, the one incorporated in our framework, is only as good as uniform weighting. So, we plan to explore and experiment on more strategies on loss weighting.

Secondly, we will research on new regularization methods, including KD and non-KD, to address USG label noise. For example, the momentum KD is another KD regularization that is widely used [12]. We have already experimented this method on MIMIC-4 dataset [11] and have seen 5 percent improvement [**?**]. In our future study, we would like to achieve a significant improvement on USG task.

Lastly, to validate the effectiveness of our method, we will evaluate the model on datasets and tasks in similar domains. For example, MIMIC-4 dataset provides healthcare data for over 40,000 patients to intensive care units, including many poten-

Table 6.1: Performance of in-hospital mortality with and without momentum knowledge distillation in MIMIC-4 dataset

model	AUROC
baseline	0.52
momentum KD	0.57

tial medical prediction tasks, such as mortality and hospital readmission, and dozens of vital measurements with rich modalities [11]. We would use datasets like MIMIC-4 to validate our MTL and KD regularization methods.

Bibliography

[1] Multiclass receiver operating characteristic (roc). https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html. Accessed: 2023-04-04.

[2] Knowledge distillation : Simplified. https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764. Accessed: 2023-04-03.

[3] George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning, 2020.

[4] Jordan J. Bird, Diego R. Faria, Cristiano Premebida, Anikó Ekárt, and George Vogiatzis. Look and listen: A multi-modality late fusion approach to scene classification for autonomous machines, 2020.

[5] R Caruana. Multitask learning. *Machine Learning*, 1997.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

[7] Changyu Deng, Xunbi Ji, Colton Rainey, Jianyu Zhang, and Wei Lu. Integrating machine learning with human knowledge. *iScience*, 23(11):101656, 2020. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2020.101656. URL https://www.sciencedirect.com/science/article/pii/S2589004220308488.

[8] Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6, 2020. doi: 10.23919/FUSION45008.2020.9190246.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[10] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa, 2019.

[11] Bulgarelli L. Pollard T. Horng S. Celi L. A. Mark R. Johnson, A. Mimic-iv (version 2.2), 2023.

[12] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021.

[13] Rongmei Lin and Xiong Li. Learning from multi-modal medical sensor data with transformer, 2022.

[14] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention, 2018.

[15] G.; Luber and M. McGeehin. Climate change and ex- treme heat events. *American journal of preventive medicine*, 2008.

[16] APRN FNP-C; Elon Lisa MA; Mix Jacqueline PhD MPH; Tovar-Aguilar Antonio PhD; Flocks Joan MA JD; Economos Eugenia; Hertzberg Vicki PhD; McCauley Linda PhD RN. Mac, Valerie PhD. Risk factors for reaching core body tem-

perature thresholds in florida agricultural workers. *Journal of Occupational and Environmental Medicine*, 2021. doi: 10.1097/JOM.0000000000002150.

[17] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning, 2020.

[18] Sebastian Ruder. An overview of multi-task learning in deep neural networks, 2017.

[19] Nocera L Shahabi C Xiong L. Tran L, Li Y. Multifusionnet: Atrial fibrillation detection with deep neural networks. *AMIA Jt Summits Transl Sci Proc*, 2020.

[20] Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. 2019.