Title

Machine Learning Methods for Personalized Healthcare

Author

Karkkainen, Kimmo

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Machine Learning Methods for Personalized Healthcare

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Kimmo Karkkainen

2023

ABSTRACT OF THE DISSERTATION

Machine Learning Methods for Personalized Healthcare

by

Kimmo Karkkainen
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2023
Professor Majid Sarrafzadeh, Chair

The escalating cost of healthcare and the growing prevalence of chronic diseases have created an urgent need for new solutions. Machine learning has the potential to revolutionize healthcare by providing more personalized and efficient care. However, there are unique challenges associated with applying machine learning in healthcare. Privacy concerns prevent data sharing across institutions, which limits available training data, and collecting individual features for patients may be invasive or expensive, as they may involve lab tests or medical imaging. In addition, machine learning models must be explainable so that medical professionals can understand how they arrive at a certain diagnosis. Despite these challenges, machine learning presents new opportunities in healthcare, both in hospitals and in remote health monitoring. In hospitals, machine learning can improve efficiency by assisting medical professionals with patient diagnoses, while in remote health monitoring, the vast quantities of data from personal and wearable devices open new opportunities for preventative care. However, processing and extracting meaningful insights from healthcare data require novel techniques. This dissertation investigates solutions for personalized healthcare in both hospital and remote healthcare settings, including adaptive data acquisition, unsupervised medical image segmentation, and remote health monitoring algorithms and applications. Overall, these solutions have the potential to improve patient outcomes and reduce healthcare costs.

The dissertation of Kimmo Karkkainen is approved.

Jungseock Joo

Cho-Jui Hsieh

Guy Van den Broeck

Majid Sarrafzadeh, Committee Chair

University of California, Los Angeles

2023

To Ela ...

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to thank my advisor, Majid Sarrafzadeh, for his guidance and support during my time in the Ph.D. program and for giving me the opportunity to work on problems that I found interesting. I am also grateful to my committee members, Jungseock Joo, Guy Van den Broeck, and Cho-Jui Hsieh, for their constructive feedback and advice. I want to acknowledge the members of the eHealth and Data Analytics Research Lab, who provided a stimulating environment and helped me develop my research skills. I also want to thank my collaborators outside the lab, both in other departments as well as in industry. Lastly, I want to acknowledge the unwavering support of my partner, whose encouragement kept me going even when there was no end in sight.

VITA

2010 – 2016 B.Sc. and M.Sc. in Computer Science and Engineering, Aalto University.

2017 – 2023 Ph.D. Student in Computer Science, University of California, Los Angeles.

PUBLICATIONS

K. Vodrahalli, G. Lyng, B. L. Hill, K. Karkkainen, J. Hertzberg, J. Zou, E. Halperin. Understanding and Predicting the Effect of Environmental Factors on People With Type 2 Diabetes. Conference on Health, Inference, and Learning (CHIL), 2023.

C. J. Cascalheira, C. Hong, R. J. Beltran, K. Karkkainen, M. Beygzade, M. Sarrafzadeh, I. W. Holloway. Analysis of Smartphone Text Data Related to Monkeypox From a Sample of Gay, Bisexual and Other Men Who Have Sex With Men. LGBT Health, 2023.

K. Karkkainen, G. Lyng, B.L. Hill, K. Vodrahalli, J. Hertzberg, E. Halperin. Sleep and Activity Prediction for Type 2 Diabetes Management Using Continuous Glucose Monitors. Learning from Time Series for Health (TS4H) Workshop at NeurIPS, 2022.

K. Karkkainen, S. Fazeli, M. Sarrafzadeh. Unsupervised Acute Intracranial Hemorrhage Segmentation With Mixture Models. IEEE International Conference on Healthcare Informatics (ICHI), arXiv preprint arXiv:2105.05891, 2021.

K. Karkkainen, J. Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age Groups. Winter Conference on Applications of Computer Vision (WACV), arXiv preprint arXiv:1908.04913, 2021.

L. Levine, M. Gwak, K. Karkkainen, S. Fazeli, B. Zadeh, T. Peris, A. Young, M. Sarrafzadeh. Anxiety Detection Leveraging Mobile Passive Sensing. EAI International Conference of Body Area Networks (Bodynets), arXiv preprint arXiv:2008.03810, 2020.

J. Joo, K. Karkkainen. Gender Slopes: Counterfactual Fairness for Computer Vision Models by Attribute Manipulation. Fairness, Accountability, Transparency and Ethics in Multimedia (FATE/MM) Workshop at ACM Multimedia, arXiv preprint arXiv:2005.10430, 2020.

O. Goldstein, M. Kachuee, K. Karkkainen, M. Sarrafzadeh. Target-Focused Feature Selection Using Uncertainty Measurements in Healthcare Data. ACM Transactions on Computing for Healthcare, arXiv preprint arXiv:1909.06772, 2020.

M. Kachuee, K. Karkkainen, O. Goldstein, S. Darabi, M. Sarrafzadeh. Generative Imputation and Stochastic Prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence, arXiv preprint arXiv:1905.09340, 2020.

K. Karkkainen, M. Kachuee, O. Goldstein, M. Sarrafzadeh. Cost-Sensitive Feature-Value Acquisition Using Feature Relevance. arXiv preprint arXiv:1912.08281, 2019.

M. Kachuee, O. Goldstein, K. Karkkainen, S. Darabi, M. Sarrafzadeh. Opportunistic Learning: Budgeted Cost-Sensitive Learning from Data Streams. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1901.00243, 2019.

CHAPTER 1

Introduction

Healthcare expenses in the United States reached 4.3 trillion dollars in 2021 (18.3% of the gross domestic product, GDP), and these costs have been rising at a faster rate than GDP since 1970 [37]. While approximately half of the increase can be attributed to rising prices of services and medications, other factors such as aging population and increasing prevalence of chronic diseases also play a major role [59]. For instance, over half of American adults have been diagnosed with a chronic disease, such as diabetes, hypertension, or cancer, with 27% of American adults having multiple chronic diseases [20]. To keep healthcare expenses manageable, it is necessary to improve preventative care and lower the cost of treating patients.

Advances in artificial intelligence (AI) and machine learning (ML) have already transformed many traditional industries. However, healthcare field has seen a slower adoption of new technologies due to the unique challenges it poses. For example, privacy concerns make obtaining training data challenging, which necessitates the development of novel unsupervised and semi-supervised machine learning models that can be trained using limited data. Furthermore, healthcare datasets are often highly imbalanced because of how rare certain diseases are. Additionally, the approval process of medical technologies is slow, sometimes taking multiple years, as extensive evidence of safety and efficacy is needed [32]. Lastly, many machine learning algorithms are difficult to interpret due to the black box nature of them, which makes it challenging to determine how the prediction was made. This lack of transparency makes clinicians less likely to choose a treatment or prescribe a drug recom-

mended by these algorithms if they do not understand what information the prediction was based on. This highlights the importance of using interpretable algorithms in healthcare.

Despite the challenges, researchers have proposed various machine learning models to address a multitude of healthcare problems. In a hospital setting, machine learning can be utilized to analyze medical images [10, 29, 47, 86, 88, 109, 153], predict patient outcomes [40, 51, 101, 176, 186], diagnose diseases [2, 5, 147, 173], or predict readmissions [73, 123, 151]. Machine learning can also aid in better utilizing patient's medical records, as the process of manually reviewing them can be time-consuming for clinicians. By summarizing the patient's clinical notes [161], machine learning can help clinicians make more informed decisions.

Machine learning can also provide benefits outside of the clinical setting, as many individuals carry a variety of sensors with them throughout the day. Modern smartphones are equipped with sensors such as microphones, light sensors, global positioning systems (GPS), and accelerometers, while smart watches, smart rings, and fitness trackers provide additional physiological signals, such as heart rate, respiration, and temperature. By analyzing sensor data from various sources, it is possible to gain a comprehensive understanding of an individual's daily habits and potential risks. This can be used to track a wide range of health problems, such as depression and anxiety [125, 134], Parkinson's disease [179], or schizophrenia [183]. Additionally, medical device manufacturers have developed specialized wearable devices for tracking specific diseases. For example, people with diabetes may wear a continuous glucose monitor (CGM) to keep track of their blood glucose levels, while wearable electrocardiograms (ECG) can be worn to identify heart problems, such as atrial fibrillation, and wearable blood pressure monitors can inform how blood pressure is changing throughout the day. All of this data can be used to provide highly personalized care to patients.

This dissertation presents a range of solutions for personalized healthcare in various settings. The first part of the dissertation focuses on improving the efficiency of hospital care. In Chapter 2, we propose an algorithm that can determine which diagnostic tests should be performed based on the currently known information, while taking into account the

varying cost of acquiring different features. Chapter 3 proposes an algorithm for unsupervised intracranial hemorrhage segmentation, addressing the challenge of segmenting images when training data is difficult to acquire due to privacy concerns. The remaining sections focus on remote health monitoring that uses data from personal or wearable devices. In Chapter 4, we introduce a novel neural network architecture that can use continuous glucose monitoring data, along with an individual's medical history, to determine daily activities. This can be used to track adherence to sleep and physical activity guidelines for individuals with Type 2 diabetes and enable new intervention and diabetes management techniques. In Chapter 5, we present a mobile sensing application that collects behavioral data to identify individuals at high risk of substance use or risky sexual behaviors. This information can be used to determine which individuals may benefit from an intervention. Together, these algorithms and applications have the potential to improve healthcare comprehensively both inside hospitals and in remote healthcare setting.

CHAPTER 2

Cost-Sensitive Feature-Value Acquisition Using Feature Relevance

2.1 Introduction

Traditionally, research on machine learning algorithms has focused on achieving accurate predictions on fully available feature sets. However, in the real world, data is often incomplete and acquiring additional feature values will incur a cost. For example, when making a medical diagnosis, a doctor examines the patient and determines what further information is needed to make a diagnosis. The next question to ask has to be chosen out of a vast number of possibilities, but the number of potentially useful follow-up questions can be narrowed down with the answers received for each question. Once the doctor has acquired enough information on the patient, they make a diagnosis and choose the appropriate treatment. Acquiring information could mean, for example, performing medical tests (blood tests, imaging, etc.) or asking the patient for more subjective information on their symptoms. In this case, there are monetary costs for each test as well as for the doctor's time, and these costs can be very different. Asking the patient for information is fast and cheap, whereas performing medical tests can have a high cost.

Money is not the only type of cost that needs to be taken into account. Medical tests can have negative side-effects or they can cause patient discomfort. In the field of mobile health, acquiring data can have an impact on the mobile device's battery life. Individuals might also be uncomfortable disclosing specific information, in which case there is a privacy cost

for acquiring data. There could even be a cost based on how much human time is required to acquire a feature value. The approaches proposed in this chapter can be used with any type of cost, as long as the costs are quantifiable on a linear scale.

There is a wide range of existing research on minimizing the costs associated with machine learning. We categorize these approaches into four different categories. The first category is feature selection, where the number of features is minimized in the training phase so that the cost becomes lower without affecting the prediction accuracy too much [39]. The second category is cascade algorithms, where at each step a decision is made to either acquire the next feature or stop and make a prediction [42, 175, 181, 188]. The third category is trees of classifiers, in which the new information affects which feature is acquired next [104, 182, 187], and the last category is fully adaptive algorithms, which can choose any feature based on what is most valuable in the current situation [43, 63, 97, 98]. There are only a few algorithms in the last category, and they tend to have a high computational cost.

The algorithms proposed in this chapter belong to the last category, where any unknown feature can be chosen for acquisition. Furthermore, the feature acquisition decisions are based on the knowledge of the existing features. We have chosen to use neural networks as our predictive model, as they have been shown to perform well across many domains. We use generic neural network architectures to demonstrate that our proposed algorithms do not depend on domain-specific structures. To choose which features to acquire next, we derive our approach from the recent research on relevance propagation [12, 124]. The focus of that line of research has been to explain why a neural network made a specific prediction, but we show that a similar approach can give us valuable information on the importance of missing feature values as well. To the best of our knowledge, relevance propagation has not been used in active feature acquisition previously. The benefit of our approach is the ease of training a model and the relatively low computational cost in both the training and testing phases.

In this chapter, we propose two algorithms for active feature acquisition. The first algo-

rithm achieves high accuracies at a low computational cost. The second algorithm improves the results of the first one at the expense of a slightly higher computational cost. Our main contributions can be summarized as follows:

- We demonstrate how relevance propagation can be utilized for the active feature acquisition problem.
- We develop an efficient algorithm that uses feature relevance information for feature acquisition. We show through experiments that feature relevances can give us valuable information on the usefulness of missing features with a very low computational cost.
- Based on the first algorithm, we develop a second algorithm that provides better results at a slightly increased computational cost.
- We compare our results with other state-of-the-art algorithms and show that our approach achieves a high accuracy with a lower cost on three realistic datasets.

2.2 Related Work

Feature value acquisition cost can be reduced in multiple ways. The most straightforward approach is to select a subset of features before or during the model training phase. This means choosing the most valuable (i.e. low cost and high informativeness) features and acquiring only them when making predictions. As there is such a wide range of feature selection algorithms, we will not describe them in detail, but an interested reader can find a survey on them from e.g. [39]. Some recent approaches can perform cost-sensitive feature selection implicitly by taking feature cost into account when training the model [129, 130]. The challenge of performing feature selection in the training phase is the lack of adaptability. These approaches acquire the same set of features for every input even though the acquired feature values could give information on which other features are useful. For some inputs, only a fraction of the features is enough to make an accurate prediction, whereas more

complicated inputs need more features. The benefit of these approaches, however, is that all the necessary computation is performed in the training phase, which makes the prediction phase computationally fast.

One approach for minimizing the feature acquisition cost for each input individually is to build a cascade of classifiers. In this approach, the algorithms have two options: make a prediction with the current classifier or request more features and move to the next classifier. With more features available, each classifier can improve on the prediction accuracy of the previous classifier. This approach has been used by e.g. [42, 175, 181, 188]. In cascade algorithms, the number of classifiers depends on the number of features, which makes it computationally expensive if the dataset contains a large number of features. It is possible to reduce the computational cost by requesting multiple features at once, but it will increase the cost of moving to the next classifier.

A more dynamic approach is to replace the static feature order with a tree of classifiers [104, 187]. In this approach, each node of a tree represents a classifier. Each child node will require additional features to improve the prediction, so moving further down the tree increases the total cost. At each node, a decision is made to either make a prediction with the current knowledge or to move down to one of the child nodes. The benefit of this approach is that the selected set of features can be different for each input. Again, a large number of classifiers is needed, which makes the training-time computational cost high when there is a large number of features. Similar to classification trees, [182] use a directed acyclic graph where each node is a classifier using a subset of features, and there are edges to other classifier nodes that gradually add more features to the requested feature set. This approach allows sharing the same classifiers in multiple paths but still requires training a large number of classifiers.

A number of approaches have been proposed to make the feature acquisition even more flexible. One of them is to try different values for each missing feature to see how much they can change the prediction. An example of this approach is FOCUS [63] which defines

expected prediction utility of each feature and chooses the feature with the highest utility, penalized by the cost of acquiring that feature. Another approach is to use Bayesian networks, which make predictions with partially known feature sets easy. [48] presented an idea of same-decision probability, which determines the probability of prediction staying the same with additional known features. This idea has been used for feature selection in e.g. [43, 50]. While these approaches provide good results, the computational complexity of choosing a feature makes them intractable for many real-world datasets with a large number of features. Recently, using sensitivity analysis on neural networks [97] and using deep reinforcement learning [90, 98] have also been proposed for feature acquisition.

Lastly, [128] proposed a different type of an approach for reducing the feature acquisition cost. Their algorithm uses two classifiers with a gating function to reduce the average prediction cost. First of these classifiers uses a small subset of the features and it is used only for the easy inputs. The second classifier can make accurate predictions for the more complicated inputs but it has a higher cost. They train the low-cost model together with a gating function, which determines when it is appropriate to use the high-cost model. This approach has been shown to have good results with a very low average cost.

2.3 Relevance Propagation

The approach proposed in this section is related to the recent research on layer-wise relevance propagation (LRP) [12]. The goal of LRP is to show how large an impact each of the input features had on the prediction, which is a non-trivial task on neural networks due to their non-linearities. To simplify this problem, relevances are propagated back to the input layer one layer at a time following a few specific rules. First, the relevance of the output layer is set to the predicted output value:

$$R_{out} = f(\mathbf{x}) \,, \tag{2.1}$$

where $f(\mathbf{x})$ is the output of our neural network with an input vector $\mathbf{x}$.

Second, each layer should also have the same total relevance:

$$R_{out} = \sum_{i=1}^{N_l} R_i^l = \sum_{i=1}^{N_{l-1}} R_i^{l-1} = \ldots = \sum_{i=1}^{N_1} R_i^1 \,, \tag{2.2}$$

where R_i^l is the output relevance of neuron i belonging to layer l and N_l is the number of neurons in layer l. This constraint guarantees that all of the relevance on the higher layer will be distributed to the neurons on the lower layer without adding or removing any relevance between layers. However, the distribution of the relevance within the layers can differ. Doing this at every layer means that the total relevance on the input layer is equal to the relevance on the output layer. Furthermore, the output relevance of a neuron should be equal to the sum of relevances of lower-level neurons that are directly connected to it:

$$R_i = \sum_j R_{i \leftarrow j} \,, \tag{2.3}$$

where $R_{i \leftarrow j}$ is the amount of relevance attributed from a higher-level neuron j to neuron i. These rules are fulfilled when a neuron distributes its output relevance entirely to its inputs according to specific rules.

There are multiple approaches for distributing the relevances to the lower layer. The approach proposed by [12] is to distribute the relevance of one neuron to the input neurons in the same proportion as the input values received from each neuron:

$$R_{i \leftarrow j} = \frac{z_{ij}}{\sum_{j'} z_{ij'}} \cdot R_j \,, \tag{2.4}$$

in which $z_{ij} = x_i w_{ij}$, x_i is an input value for the neuron, and w_{ij} in the corresponding weight. As this rule might not be numerically stable with small activation values, they present a few alternative approaches to mitigate the problem, but the overall idea remains the same.

[124] proposed another way to propagate relevances while still following the other rules defined by [12]. Their approach, called Deep Taylor Decomposition, approximates each neuron using a first-order Taylor series approximation to determine how much relevance should be propagated to each neuron input. Using this approach, they derive three rules for propagating the relevance based on the neuron's input domain. When the neuron has unconstrained input, the propagation rule becomes:

$$R_i = \sum_j \frac{w_{ij}^2}{\sum_{i'} w_{i'j}^2} R_j \,. \tag{2.5}$$

If the input is constrained to have only positive values, as is the case when the previous layer contains rectified linear units (ReLU), the propagation rule is:

$$R_i = \sum_j \frac{z_{ij}^+}{\sum_{i'} z_{i'j}^+} R_j \,, \tag{2.6}$$

where $z_{ij}^+ = x_i w_{ij}^+$, and w_{ij}^+ is the positive part of the weight w_{ij}. The positive part of a vector is defined as a vector that has all the negative values replaced by zeroes. Finally, if the neuron input is known to have specific upper (h_i) and lower (l_i) bounds, as is often the case in the first layer, the rule becomes:

$$R_i = \sum_j \frac{z_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_{i'} z_{i'j} - l_{i'} w_{i'j}^+ - h_{i'} w_{i'j}^-} R_j \,. \tag{2.7}$$

Full derivations of these rules can be found in [124].

2.4 Our Approach

2.4.1 Problem Definition

We start by considering a case where a feature vector $\mathbf{x}_{full}$ has all feature values available, but only part of these values are known to us at any time. We know the value of $\mathbf{x}_{partial} =$

$\mathbf{x}_{full} \odot \mathbf{k}^t$, where $\mathbf{k}^t$ is an indicator vector containing values 1 or 0 depending on whether the corresponding feature value is known to us at time t. The number of known features at time t is therefore $|\mathbf{k}^t|_0$. To acquire a new feature, we have to pay feature acquisition cost c_i, which can be different for each feature i.

The feature acquisition algorithm will then take the following steps. First, the algorithm should determine the most valuable feature using the knowledge of the acquired feature values, the current prediction, and feature costs. This feature value is then requested from an oracle that is able to give any value for a cost, and the value is added to the feature vector of known values $\mathbf{x}_{partial}$. The cumulative cost at each time step is therefore $\sum_i c_i k_i^t$. This process is repeated until a stopping condition is reached. Possible stopping conditions can include, for example, the model certainty reaching a predefined level, the total cost becoming too high, or all of the feature values having been acquired. The appropriate stopping condition should be decided based on the problem domain. For example, in the medical domain it might be more appropriate to stop only when the model certainty is high enough, even if it leads to a higher cost. In less critical domains, minimizing the total cost could be more important.

An outline for a generic, incremental feature acquisition algorithm is shown in Algorithm 1. The goal of this algorithm is to acquire feature values until an accurate prediction can be made while minimizing the cost. This algorithm assumes that it is possible to acquire each missing feature for a cost. This requirement can be relaxed by skipping the highest scoring feature values that can not be requested (e.g. the cost is prohibitively high or a specific test cannot be performed in that particular situation).

Our goal is then to define a value function $value(\mathbf{x}_{\text{partial}}, \hat{\mathbf{y}}, \mathbf{c})$, which uses the currently known features, the current prediction, and the feature costs to determine the value of acquiring each unknown feature. This function should give the highest value for features that increase the prediction accuracy the most while simultaneously having a low cost.

In this section, we propose two methods for cost-sensitive feature acquisition where rele-

Algorithm 1: Outline for the feature acquisition algorithm

Input: Feature vector $\mathbf{x}_{partial}$, model M, cost vector $\mathbf{c}$

repeat

$t \leftarrow t + 1$

$\hat{\mathbf{y}} \leftarrow M.predict(\mathbf{x})$

$\mathbf{v} \leftarrow value(\mathbf{x}_{partial}, \hat{\mathbf{y}}, \mathbf{c})$

$next\_feature \leftarrow argmax_i(\mathbf{v})$

$\mathbf{x}^{t+1}_{partial} \leftarrow acquire(\mathbf{x}^{t}_{partial}, next\_feature)$

until Stopping condition is reached

return $M.predict(\mathbf{x})$

vance propagation is used to determine feature informativeness. On a high level, the goal of these algorithms is to define which unknown feature is the most important for our current prediction at any given time. This importance should reflect our current knowledge of the available feature values. Our first method propagates the predicted value directly similar to how relevance propagation has been used to determine features' impact on prediction in prior works. Our second method propagates relevance through each output node separately to determine which feature is the most important over any potential class.

2.4.2 Direct Propagation

We start by training a neural network model M with fully available features. Having fully available features is not a strict requirement but we choose to use them to make the results comparable with the existing research in this field. The model M takes an input vector $\mathbf{x} \in \mathbb{R}^m$ and produces a prediction $\hat{\mathbf{y}} \in \mathbb{R}^n$. Our algorithm receives an input vector $\mathbf{x}_{partial}$, which has no known values initially. To keep track of which feature-values have been acquired already, we introduce an indicator vector $\mathbf{k} \in \mathbb{R}^m$, which has value 1 if the corresponding feature-value has been acquired and value 0 otherwise. The goal is then to request these

feature values one-by-one until a prediction can be made. First, we fill in the missing values of the given input vector $\mathbf{x}_{partial}$:

$$\mathbf{x}_{filled} \leftarrow \mathbf{k} \odot \mathbf{x}_{partial} + (\mathbf{1} - \mathbf{k}) \odot \mathbb{E}[\mathbf{x}]\,, \tag{2.8}$$

where $\odot$ is the Hadamard product. The true expected value of $\mathbf{x}$ is not known, so we estimate it using the training data. This gives us a feature vector where the missing values have been replaced with the expectation for those values. We then propagate $\mathbf{x}_{filled}$ forward through our model M to acquire a prediction $\hat{\mathbf{y}}$. This prediction is then used as the starting point for relevance propagation. Using the Deep Taylor Decomposition propagation rules described in Section 2.3, $\hat{\mathbf{y}}$ is propagated backward to the input nodes. This gives us relevances r_i for each input node, which is demonstrated in Figure 2.1.

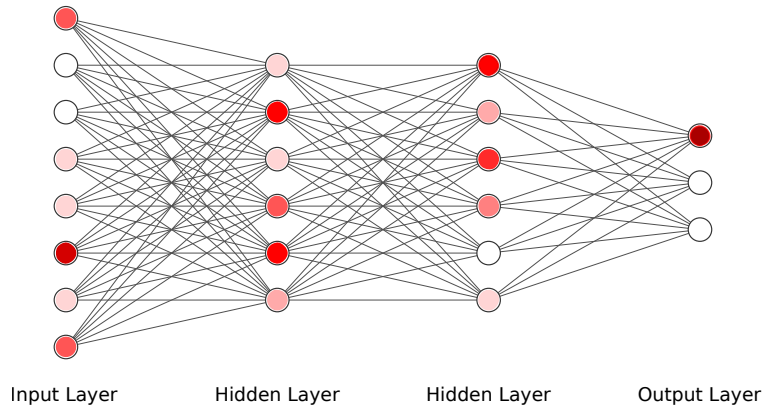


Figure 2.1: Relevance is propagated from the output layer (on the right) to the input layer (on the left). Darker color indicates a higher relevance.

Next, we will introduce an adjusted relevance score to take into account the constraints that our problem has. First, we need to notice that the relevances can have either positive or negative values, depending on whether the corresponding feature increased or decreased the predicted value. High impact in either direction can be important, so we will use the absolute value. Next, we need to take into account the feature costs. Two features could provide the same amount of information but have a vastly different cost, so we will normalize

the earlier value using the corresponding feature cost. Finally, we are only interested in the unknown features, which leads us to the adjusted relevance $\hat{\mathbf{r}}$ where individual values are defined as:

$$\hat{r}_i \leftarrow (1 - k_i) \cdot \frac{|r_i|}{c_i}\,. \tag{2.9}$$

This defines all known feature values to have zero relevance, whereas unknown feature values have a relevance value depending on the magnitude of their original relevance as well as the associated feature acquisition cost. The relevance of known features is set to zero to avoid acquiring the same features multiple times. Finally, we acquire the feature value x_j, where $j = argmax(\hat{\mathbf{r}})$. The direct propagation approach is shown in Algorithm 2. This approach performs one forward and one backward propagation for each acquired feature, so the computational complexity is $O(m)$.

Algorithm 2: Direct propagation algorithm

Input: Feature vector $\mathbf{x}_{partial}$, model M, cost vector $\mathbf{c}$

$t \leftarrow 0, \mathbf{k} \leftarrow \mathbf{0}$

$\mathbf{x}^t_{filled} \leftarrow \mathbf{k} \odot \mathbf{x}_{partial} + (\mathbf{1} - \mathbf{k}) \odot \mathbb{E}[\mathbf{x}]$

repeat

$\hat{\mathbf{y}} \leftarrow M.predict(\mathbf{x}^t_{filled})$

$\mathbf{r} \leftarrow abs(get\_relevances(\mathbf{x}^t_{filled}, \hat{\mathbf{y}}))$

$\hat{\mathbf{r}} \leftarrow (\mathbf{r} \odot (\mathbf{1} - \mathbf{k})) \oslash \mathbf{c}$

$next\_feature \leftarrow argmax_i(\hat{\mathbf{r}})$

$\mathbf{x}^{t+1}_{filled} \leftarrow acquire(\mathbf{x}^t_{filled}, next\_feature)$

$k_{next\_feature} \leftarrow 1$

$t \leftarrow t + 1$

until Stopping condition is reached

return $M.predict(\mathbf{x}^t_{filled})$

2.4.3 Multiple Propagations

Our second algorithm modifies the direct propagation algorithm to take into account relevant features for each output node $\hat{y}_i$. The intuition behind this approach is that some output nodes might receive a low predicted value due to some unknown input feature, which can cause that input feature to have a low relevance in the direct propagation algorithm. By propagating relevance through each output node separately, we can avoid such a situation at the cost of increased computational complexity.

The problem setting is the same as in our first algorithm. However, when the relevance was propagated backward in the previous method, this time we set $\hat{y}_i = 1$ for one i, while $\hat{y}_j = 0, \forall j \neq i$. This is then propagated backward using the same rules as previously. This process is repeated for all values of i, so in total the backward propagation is performed n times, once for each output class. We then choose the globally maximal adjusted relevance $\hat{r}_i$, and acquire the value of feature i.

The multiple propagations method is shown in Algorithm 3. This method is computationally more demanding than the first method, as each feature acquisition requires one forward pass and n backward passes, thereby making the computational complexity $O(nm)$. However, the algorithm can be modified to perform the backward passes in parallel, as they do not depend on each other. For clarity, only the serial method is shown here.

2.4.4 Implementation details

We implement our algorithm using a fully-connected neural network, where the number of layers and number of neurons, as well as other hyperparameters, are chosen by optimizing the network to have as high accuracy as possible on the fully available training data. We use PyTorch [140] for our implementation, in which we have extended the necessary neural network layers to support relevance propagation. In addition, we use as high L2 regularization as possible without decreasing the validation accuracy. Based on our empirical observations,

Algorithm 3: Multiple propagations algorithm

Input: Feature vector $\mathbf{x}_{partial}$, model M, cost vector $\mathbf{c}$

$t \leftarrow 0, \mathbf{k} \leftarrow \mathbf{0}$

$\mathbf{x}_{filled} \leftarrow \mathbf{k} \odot \mathbf{x}_{partial} + (\mathbf{1} - \mathbf{k}) \odot \mathbb{E}[\mathbf{x}]$

repeat

  $best\_relevance \leftarrow -\infty$

  $best\_feature \leftarrow 0$

  for $i \leftarrow 0 \ldots n\_classes$ **do**

    relevance_out $\leftarrow \mathbf{0}$

    $relevance\_out_i \leftarrow 1$

    $\mathbf{r} \leftarrow get\_relevances(\mathbf{x}^t_{filled}, \mathbf{relevance\_out})$

    $\hat{\mathbf{r}} \leftarrow (abs(\mathbf{r}) \odot (\mathbf{1} - \mathbf{k})) \oslash \mathbf{c}$

    if $max(\hat{\mathbf{r}}) > best\_relevance$ **then**

      $best\_feature \leftarrow argmax(\hat{\mathbf{r}})$

      $best\_relevance \leftarrow max(\hat{\mathbf{r}})$

    end if

  end for

  $\mathbf{x}^{t+1}_{filled} \leftarrow acquire(\mathbf{x}^t_{filled}, best\_feature)$

  $\mathbf{k}_{best\_feature} \leftarrow 1$

  $t \leftarrow t + 1$

until Stopping condition is reached

return $M.predict(\mathbf{x}^t_{filled})$

proper regularization improves the feature acquisition results, even if it does not affect the prediction accuracy on fully available data.

2.5 Experiments and Results

We evaluate our approaches on three cost-sensitive datasets: diabetes prediction, heart disease prediction, and Yahoo! Learning to Rank Competition dataset (LTRC). Summary of these datasets can be found in Table 2.1.

Table 2.1: Dataset statistics

Dataset	**Examples**	**Features**	**Classes**
LTRC	30000	500	5
Diabetes	25474	581	3
Heart	49509	245	2

2.5.1 Diabetes Prediction

As medical diagnosis is an area where active feature acquisition can be highly beneficial, the first two experiments use medical datasets. We have derived two datasets from The National Health and Nutrition Examination Survey (NHANES), which is a long-term program that has collected health and nutrition data from a nationally representative group of 5000 people between years 1999–2016 [34]. The full dataset contains questionnaire answers as well as results from physical and laboratory examinations. As this dataset contains data on a wide range of medical problems, we will first focus on predicting whether an individual has diabetes, prediabetes, or no diabetes.

To predict the level of diabetes, we look at the blood glucose levels and define an individual as healthy if their fasting plasma glucose level is below 100 mg/dL, prediabetic if

the level is between 100 and 125 mg/dL, and diabetic if the level is over 125 mg/dL. These ranges have been defined by the Centers for Disease Control and Prevention (CDC) [33]. We filter out features that are directly related to our target variable, i.e. any variable measuring blood glucose, as they would make the prediction trivial. We also leave out variables that have missing values for over 25% of the subjects. We define feature costs based on a rough estimate on how much money and effort is needed to acquire that feature. The feature costs are listed in Table 2.2. The dataset is provided in the supplemental materials with the code to reproduce our results. The data was split into a training set (70%) and a test set (30%) randomly. Both sets were balanced by oversampling to have an equal number of examples from each class. All features were normalized to range $[0, 1]$.

Table 2.2: Feature costs for NHANES dataset

Feature type	**Cost**
Demographics	1
Questionnaire answer	5
Physical examination	10
Laboratory test	100

We compare our approach to four other algorithms: The Greedy Miser [188], AdaptApprox [128], FACT [97] and Opportunistic Learning (OL) [98]. We selected these algorithms based on their good performance and the availability of reference implementations. The Greedy Miser is an algorithm for learning cost-sensitive classification and regression trees. AdaptApprox learns a gating function to decide whether to use a cheap or an expensive classifier. FACT uses neural network's sensitivity to determine which feature is most likely to change the prediction. Opportunistic Learning uses reinforcement learning to learn what features to ask. The results of the first experiment can be found in Figure 2.2.

In this experiment, both of our proposed algorithms (Relevance-DP, Relevance-MP) provide nearly identical results. Initially, FACT provides similar results to our algorithms, but

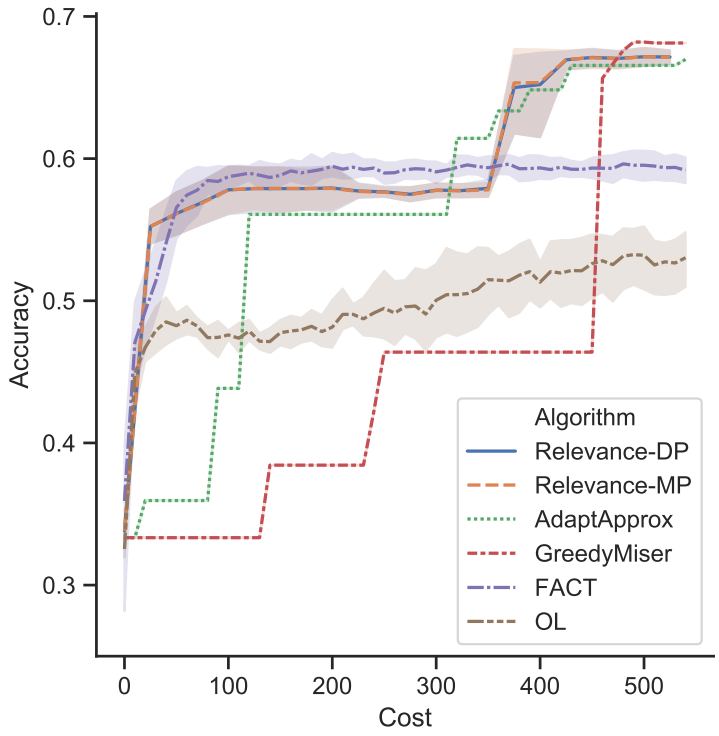


Figure 2.2: Prediction accuracy on diabetes dataset.

converges to a lower accuracy. Opportunistic Learning (OL) improves its accuracy very slowly after the initial features. AdaptApprox reaches a good accuracy but slightly slower than our algorithms. The Greedy Miser has to acquire a large number of features before finally achieving similar accuracy to the other algorithms. In this experiment, our algorithms provide a fast convergence and a high final accuracy, thereby combining the best aspects of the other algorithms.

2.5.2 Heart Disease Prediction

Our next dataset is also derived from the NHANES dataset [34]. This time the goal is to predict whether an individual has a heart disease, such as congestive heart failure, or has had a heart attack. We use the same costs and data preprocessing steps as in the previous dataset, shown in Table 2.2.

The results are shown in Figure 2.3. This time there is a significant difference between our proposed algorithms. Using multiple propagations (Relevance-MP) provides faster convergence than using direct propagation (Relevance-DP). Again, FACT provides similar results to our Relevance-MP algorithm initially but converges to a lower accuracy. Opportunistic Learning (OL) starts slightly slower and converges to a similar accuracy with FACT. Adapt-Approx starts with more expensive features, therefore staying at a low accuracy for longer. The Greedy Miser suffers from the same problem, but also reaches a good accuracy later. This experiment shows the benefit of using multiple propagations instead of the more simple algorithm.

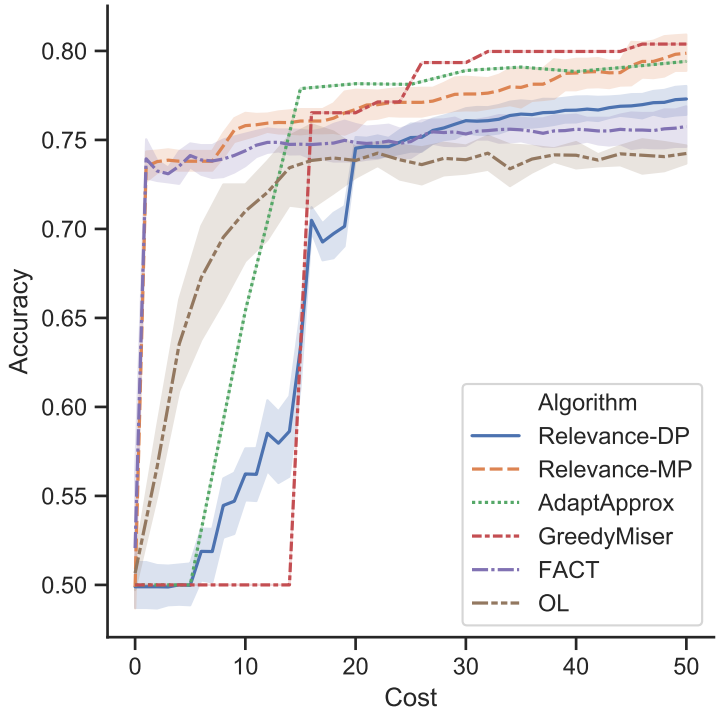


Figure 2.3: Prediction accuracy on heart disease dataset.

2.5.3 Learning to Rank Competition

The goal of Yahoo! Learning to Rank Competition (LTRC) dataset is to predict how relevant a specific document is given a query [41]. The relevance is defined by an expert using a five-step scale. The feature vectors consist of features describing the query and the document. These features include e.g. how many times a document has been clicked on the result list, how recent the page is, how well the document's text matches the query, and so on. Each feature has an associated feature extraction cost between 1–200, which has been defined by Yahoo!. The split into training and evaluation sets has been defined in the original dataset and was used as-is for our experiments.

Normalized Discounted Cumulative Gain (NDCG) has been suggested as a measure of quality for the predictions on this task [41]. This metric was introduced by [92], and it compares the relevance of the predicted result order to the optimal order. It has also been used by previous feature acquisition papers [187, 188], so we will use it to measure the performance of our algorithms as well.

As the ranking problem is different from the traditional classification problem, we compare our results to the algorithms that have been designed for this problem and have demonstrated the best performance: Cost-Sensitive Tree of Classifiers (CSTC) [187], Cronus [42] and Early Exit [27]. CSTC builds a tree of classifiers, where the chosen path determines which features will be used for prediction. Cronus builds a cascade of classifiers so that the easy inputs will be handled by the earlier classifiers in the cascade with a low cost, and the more complicated inputs will go through more classifiers leading to a higher cost. Early Exit scores documents gradually, dropping out the ones with too low scores before fully evaluating them.

The results are shown in Figure 2.4. As can be seen, feature acquisition with relevance propagation reaches a higher NDCG@5 score than the other approaches. In addition, using multiple propagations (Relevance-MP) converges to a high score significantly faster than

using the direct propagation algorithm (Relevance-DP).

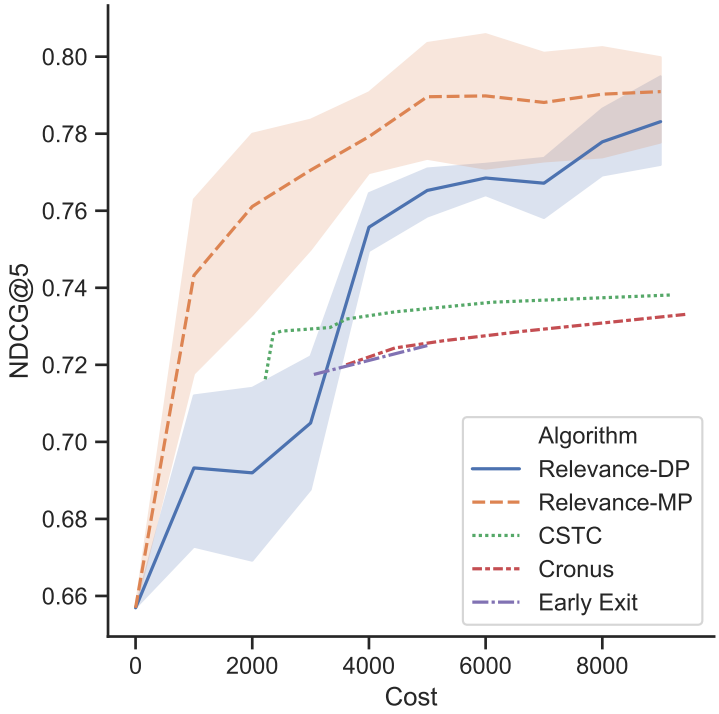


Figure 2.4: NDCG@5 score on Yahoo Learning to Rank Competition dataset.

2.6 Discussion

In this chapter, we have presented two active feature acquisition algorithms that outperform prior algorithms in terms of accuracy and cost. We compared our algorithms to various algorithms based on neural network sensitivity analysis [97], reinforcement learning [98], classification trees [187, 188], and gating algorithms [42, 128], and our multiple propagations algorithm consistently outperformed the prior algorithms. Our single propagation algorithm performed well in most cases, but it did not reach as good results with the heart disease prediction and LTRC dataset. We believe this to be due to its focus on the class with the highest predicted probability. This may lead the algorithm to focus on non-optimal features if the initial prediction is wrong. In contrast, the multiple propagations algorithm avoids

this problem by evaluating all target classes one-by-one. This, however, leads to higher computational cost. This may not be a significant problem in practice, as the backward passes can be performed independently and they are therefore easily parallelizable. Our results also demonstrate that our algorithms converge at the same or higher accuracy than prior algorithms, making our models competitive at any budget.

Our results could be further improved by using more advanced feature imputation techniques. While our algorithms used mean imputation to ensure that the evaluation focused on the performance of the feature acquisition algorithm and not on the imputation technique, this simple imputation method may not always lead to realistic imputed values. In situations where the feature value distribution is bimodal or follows some other non-normal distribution, the imputation technique could be improved by incorporating the known values. This can be done, for example, by using denoising autoencoders [180].

One limitation of our work is that the evaluation was based on retrospective datasets, so further research is needed to evaluate the algorithms' performance in a real hospital setting. It is especially important to evaluate how active feature acquisition algorithms may affect treatment when the physician disagrees with the algorithm's proposed test or diagnosis. Another limitation is that the proposed algorithms are greedy, which means that they choose a single feature with the best informativeness and cost. However, in the real world, costs and informativeness are not always independent across features. For example, performing multiple laboratory tests at the same time may be cheaper than performing them individually. In addition, making certain diagnoses may require information from multiple tests, and knowing the results of an individual test might not be highly informative by itself. This myopia may in some situations mean that the algorithm does not choose features in the globally optimal order. A simple solution is to identify feature groups that should always be acquired together. However, with more complex data, this may not be feasible. There may be a need to have overlapping feature groups, which may lead to a large number of groups. Future work is therefore still needed to explore solutions that are able to handle

feature groups efficiently.

While our research aims to improve hospital care by assisting medical personnel in decision making, these same algorithms can be applied to any situation where data acquisition is expensive. For instance, these algorithms could be used to reduce battery consumption when using machine learning on mobile phones, or they could reduce the network load when collecting data from Internet-of-Things (IoT) devices. Additionally, they could be used when processing power is the limiting factor, such as when extracting features from a high-resolution video stream.

2.7 Conclusion

In this chapter, we have presented a novel approach to the cost-sensitive active feature acquisition problem that has the potential to make hospital care more efficient by assisting medical professionals in selecting optimal tests. Our approach uses missing feature relevance as the core idea for choosing which features to acquire. We have presented two different feature acquisition algorithms using this approach. First of them provided good results with one forward and one backward propagation per acquired feature, while the second algorithm improved the results further at the expense of a slightly increased computational cost. We evaluated the proposed algorithms on three realistic datasets: Yahoo! Learning to Rank Competition dataset and two health datasets derived from the National Health and Nutrition Examination Survey (NHANES) dataset. Our results show that our first algorithm (direct propagation) performs well in most cases, while our second algorithm (multiple propagations) is more robust and out-performs the existing algorithms.

CHAPTER 3

Unsupervised Intracranial Hemorrhage Segmentation With Gaussian Mixture Models

3.1 Introduction

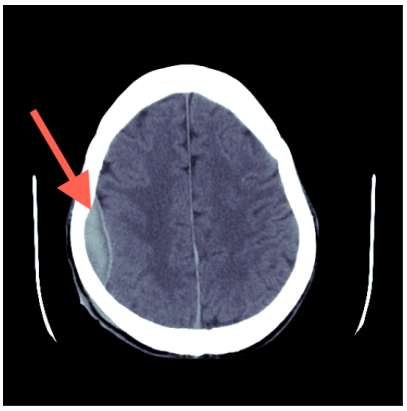
Intracranial hemorrhage is a life-threatening condition that can be caused by either physical trauma or by various medical conditions, such as high blood pressure or an aneurysm [81]. It is the cause of 15-20% of strokes, with an estimated 5 million cases per year globally [103,192]. Depending on the type of the hemorrhage, the mortality rate can be as high as 57% [78]. Approximately half of the deaths occur within the first 48 hours, which is why the treatment must be started as early as possible [24]. Before the treatment can be started, the hemorrhage must be diagnosed from medical images. This process, however, is very time-consuming due to the large number of images per patient and the complexity of the task. Therefore, automating the analysis process allows us to make the diagnosis faster which leads to faster treatment.

Intracranial hemorrhage is typically diagnosed by Computed Tomography (CT) imaging. CT produces a number of cross-sectional slices by combining information from X-ray images taken from different angles. The intensity value of the voxels indicates how much the X-ray was attenuated, and it can be used to determine the type of the tissue at each location. Attenuation is measured in Hounsfield Units (HU) which ranges from -1000 for air to 0 for water and to +2000 for dense bones. Brain matter is typically between 20-45 HU while hemorrhage usually starts with a slightly higher attenuation (e.g. 75-85 HU for subdural

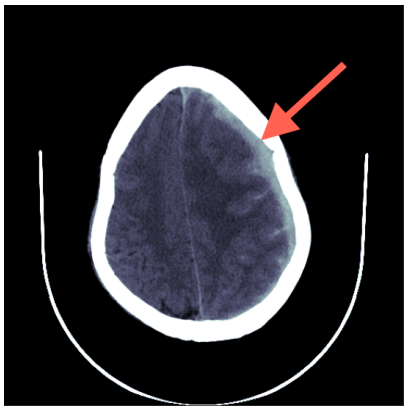
hemorrhage) but drifts closer to the attenuation levels of the healthy brain matter over time [142]. In practice, however, there can be a significant overlap in the intensity values of hemorrhage and healthy tissues. In addition, intracranial hemorrhages can differ in shapes and sizes which further complicates the analysis.

There are five categories of hemorrhages with distinct locations and shapes: subdural, epidural, subarachnoid, intraparenchymal, and intraventricular (sample images are shown in Figure 3.1). One patient can have more than one type of hemorrhage at once. Both epidural and subdural hemorrhage occur close to the skull and they can be distinguished by their shapes: epidural hemorrhage typically has a lentiform shape while subdural hemorrhage has a crescent shape. Intraparenchymal hemorrhage occurs within the brain tissue and typically has a rounded shape. Intraventricular hemorrhage is located inside the brain cavities that are otherwise shown as darker areas close to the center of the brain. Finally, subarachnoid hemorrhage occurs in the space surrounding the brain tissue. While each of these types can have different causes and symptoms, they are typically visible as lighter-than-expected regions in the scan. Other indicators include unusual shapes caused by the pressure, such as a midline that has been pushed slightly in the opposite direction from the hemorrhage.

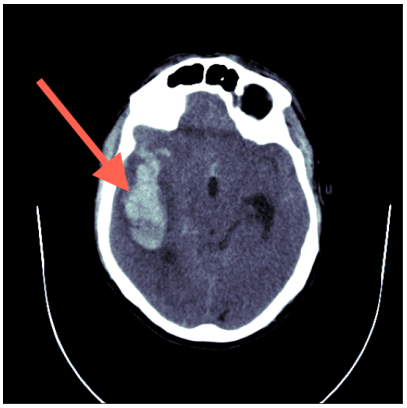
In this chapter, we propose a fully-unsupervised algorithm for acute intracranial hemorrhage segmentation based on the idea of Mixture Models. We start by demonstrating how to extract the brain from a CT scan containing various tissues and noise. Next, we show how the brain can be represented as a mixture of probability distributions where the location and intensity of different tissue types follow different probability distributions. This representation allows us to find the optimal distribution parameters using the Expectation-Maximization algorithm. We then provide an algorithm for fitting the model when the number of hemorrhage clusters is unknown. Finally, we evaluate our algorithm against other unsupervised and supervised algorithms using publicly-available datasets and provide a visual comparison of the results to demonstrate how the algorithms differ.



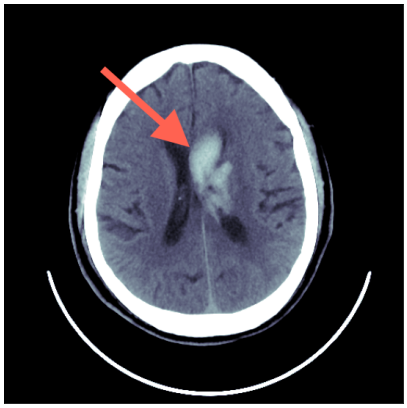
(a) Epidural



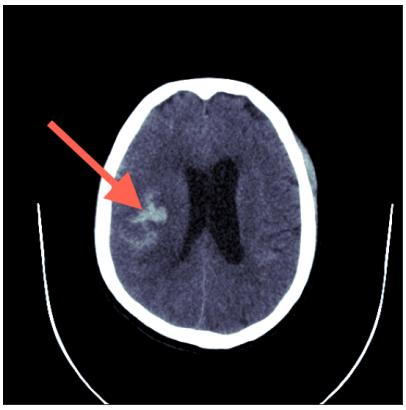
(b) Subdural



(c) Intraparenchymal



(d) Intraventricular



(e) Subarachnoid

Figure 3.1: Types of intracranial hemorrhage

3.2 Related Works

There are both unsupervised and supervised algorithms for intracranial hemorrhage segmentation. Unsupervised algorithms typically utilize the prior knowledge that hemorrhage voxels tend to have a higher intensity value than healthy voxels. The naive approach would be to determine a hard threshold and mark all voxels above that threshold as hemorrhage and all voxels below that threshold as healthy. In practice, this approach would lead to inaccurate segmentation due to the large overlap of voxel intensities between healthy tissue and hemorrhage. Depending on the chosen threshold, some low-intensity parts of the hemorrhage could be segmented as healthy while some high-intensity parts of healthy tissues could be segmented as hemorrhage. Therefore, intensity threshold alone is not sufficient for segmentation, but many algorithms use it as a starting point and refine the results with other techniques.

In earlier works, there have been many suggestions on how to determine the initial intensity threshold. Some techniques include performing Fuzzy C-Means or K-Means clustering on the intensity levels [18,19,69,115,126,152], using Otsu's method [3,167], or comparing the histogram to the expected histogram [143]. Once an appropriate threshold has been determined, the results can be refined by using active contouring [18,19,110], statistical analysis of the clusters [126], analysis of voxel neighborhoods [52], or region growing [115].

A different approach proposed by [71] is to compare the brain scan to a healthy template scan. As each brain scan is unique, the scan must first be transformed into a normalized image. This normalized image is then compared to a template and the differences are scored. Once an abnormal region is detected, the image is transformed back into the original space to be visualized.

There are also multiple supervised techniques that have been proposed for hemorrhage segmentation. For example, traditional classifiers, such as random forests or logistic regression, have been trained using hand-crafted features on voxels and their neighborhoods

[4, 127, 149]. These hand-crafted features have included, for example, voxel intensity, statistics of the voxel's neighborhood (mean, standard deviation, skew, kurtosis), percentage of high-intensity pixels in the neighborhood, difference of intensity to the opposite side of the brain, and many more. With the recent advancements in deep learning, there have also been many proposed variants of the U-Net [145] architecture [10, 47, 86, 88, 109, 153] as well as other neural network architectures [76, 89]. While supervised algorithms can provide accurate segmentations, they depend on the availability of vast quantities of segmented training data which is rarely published due to privacy concerns.

3.3 Methodology

3.3.1 Data

To develop our algorithm, we used images from a large-scale CT scan dataset published by the Radiological Society of North America (RSNA) as a part of a Kaggle competition [99]. The images in this dataset are not segmented but the dataset contains labels for each slice determining the type of the hemorrhage which allows for a visual evaluation of the results. However, the lack of ground-truth segmentations makes it unsuitable for an objective comparison of algorithms.

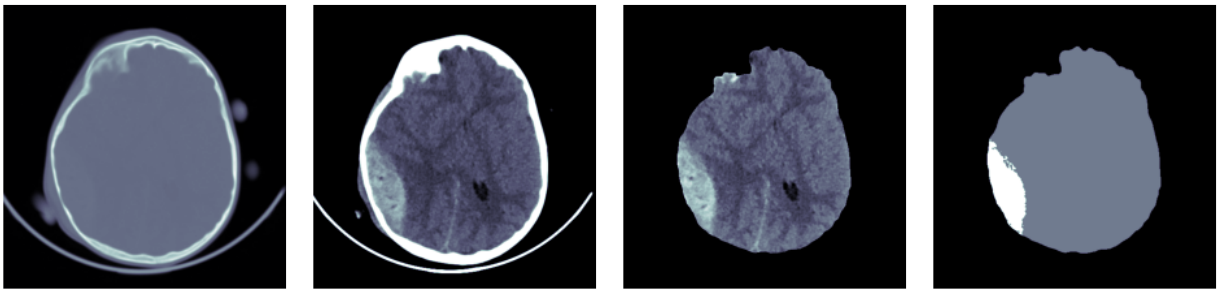
We therefore used two other publicly-available datasets for our evaluations. The first dataset has been published by Qure.ai under the Creative Commons license [46]. This dataset consists of 491 CT scans, each scan belonging to a different patient. In addition, each cross-sectional slice has been annotated by a radiologist to denote the type(s) of hemorrhage present in that slice. Out of the 491 patients, 205 patients were labeled as having some type of hemorrhage while the remaining patients were healthy. The mean age of the patients was 48.08 years, and 178 of the patients were female and 313 were male. The scans were collected using six different CT scanner models: GE BrightSpeed, GE Discovery CT750 HD, GE LightSpeed, GE Optima CT660, Philips MX 16-slice, and Philips Access-32 CT.

The annotations were provided by three senior radiologists and the ground truth label was determined by majority vote. In addition, this dataset has been developed further by [144] to include the bounding boxes of hemorrhages within each slice. This bounding box annotation was performed by three trained neuroradiologists. It should be noted that the bounding boxes only indicate where the hemorrhage is located and how large it is approximately, but it does not show us the exact outline of the hemorrhage. Our evaluation will therefore only consider the overlap between our segmention and the bounding box which might contain healthy regions as well.

The second dataset used for evaluation has been published by [85, 86] and is publicly available on PhysioNet [74]. This dataset consists of 82 CT scans with 36 of them containing hemorrhage. Regions with hemorrhage have been delineated by two radiologists who both agreed on the ground-truth segmentations. The mean patient age was 27.8 years with a standard deviation of 19.5. There were 46 male patients and 36 female patients. The CT scanner model was Siemens SOMATOM Definition Edge. As this dataset is very limited in size compared to the Qure.ai dataset but provides accurate outlines of the hemorrhages, we used it only for a visual comparison of the segmentation results.

3.3.2 Preprocessing

First, the scan is loaded from DICOM [58] or NiFTI [133] files and converted into a 3-dimensional array. An example of one slice of this array is shown in Figure 3.2a. Due to the large range of intensity values (air having very low values while bone having very high values), the small intensity variations within the brain are not visible without preprocessing and all brain tissue looks similar. Then, we perform windowing which removes the very high and very low intensity values which are known to be irrelevant to the segmentation task:



(a) Original image (b) Windowed image (c) Extracted brain (d) Segmented brain

Figure 3.2: Image processing pipeline

$$\text{intensity}_{\text{windowed}} = \begin{cases} i_{\min} & \text{, when intensity} < i_{\min}, \\ i_{\max} & \text{, when intensity} > i_{\max}, \\ \text{intensity} & \text{, otherwise} \end{cases} \tag{3.1}$$

where $i_{\min}$ and $i_{\max}$ are the window boundaries. The boundaries should be chosen s.t. we remove as much of the irrelevant variations as possible without losing any information related to the brain tissue and hemorrhage. In our experiments, we have chosen conservative limits of $i_{\min} = 0$ and $i_{\max} = 100$, which is a wide enough range to cover all of the brain matter and blood while at the same time ignoring the intensity variations of the other tissues, such as bones. It also makes visual evaluation of the images feasible, as we can focus on the minor differences in this narrow range. An example of a windowed image is shown in Figure 3.2b.

After windowing, all of the bone tissue has the same intensity, which makes it easier to remove. We start by selecting all tissues with intensity $i_{\max}$ as the removal mask. This mask might have gaps due to possible skull fractures, so we fill the small gaps by using morphological closing. The areas covered by this mask are then set to $i_{\min}$. After removing the skull, there are still some soft tissues which are not part of the brain. We remove them by finding the largest contiguous region which we assume to be the brain. We set all regions that are not connected to the largest contiguous region to $i_{\min}$. In addition, a narrow slice

along the edge of the brain is removed by using morphological erosion, as it often has a higher intensity and might be mistakenly detected as hemorrhage. An example of a fully preprocessed image can be seen in Figure 3.2c.

3.3.3 Mixture Model

After all non-brain voxels have been removed from the scan, our goal is to determine which voxels correspond to healthy tissues and which voxels correspond to hemorrhage. Our model makes the assumption that different tissue types follow different distributions and therefore we can represent the brain scan using a mixture model. This mixture model should take into account both the location and the intensity of the voxel. For example, a high intensity value could be caused by either hemorrhage or noise depending on whether it is surrounded by a larger region of high-intensity voxels or not. We also make the assumption that the hemorrhage voxels follow a Gaussian distribution in both the coordinate and the intensity space, i.e. the hemorrhage voxels are located close together and have similar intensity values. However, healthy voxels are spread evenly throughout the brain, so we assume that their intensity values follow a Gaussian distribution but the location follows a Uniform distribution. This approach differs from the earlier algorithms that start by considering only the intensity values and then refine the results by taking coordinates into account as a secondary step. Our algorithm is able to optimize the results in both spaces simultaneously.

Similar intensity values typically represent similar tissue types regardless of the location, so we start with the assumption that the voxel intensity is independent of the voxel location:

$$P(int, x, y, z|c) = P(int|c)P(x, y, z|c) \tag{3.2}$$

Here, int is the intensity of a voxel in coordinates (x, y, z) and c is the cluster index. For convenience, we define that the first cluster always corresponds to the healthy tissue and there might be additional clusters that correspond to hemorrhage. We also make the

assumption that the intensity of each cluster c follows a Gaussian distribution:

$$P(int|c) \sim \mathcal{N}(\mu_c, \sigma_c), \tag{3.3}$$

where μ_c and σ_c are the mean and standard deviation of the intensity values of cluster c respectively. If there are multiple hemorrhage clusters, this allows them to have different intensity distributions, which is necessary because the intensity values could differ based on the type and age of the hemorrhage. As mentioned earlier, we make the assumption that the location of healthy voxels follows a uniform distribution while the location of hemorrhage voxels follows a Gaussian distribution:

$$P(x, y, z|c) \sim \begin{cases} \mathcal{U}(0, N_{\text{bv}}) & \text{, when } c = 0 \\ \mathcal{N}(\mu_{\mathbf{loc,c}}, \mathbf{\Sigma}_{\text{loc,c}}) & \text{, when } c \geq 1 \end{cases} \tag{3.4}$$

N_{bv} is the number of brain voxels (excluding skull and outside areas), $\mu_{\mathbf{loc,c}}$ is the center of cluster c, and $\mathbf{\Sigma}_{\text{loc,c}}$ is the covariance matrix of the location. To determine the optimal parameter values, we use the Expectation Maximization (EM) algorithm [56], which is an iterative approach for finding the missing parameters by alternating between calculating the expected cluster membership values (E-step) and calculating the missing parameters (M-step). Using the previously defined distributions, the update rule for the cluster memberships (E-step) becomes:

$$\begin{aligned} \gamma_{i,c} &= \frac{\pi_c p(i|c)}{p(i)} \\ &= \begin{cases} \frac{\pi_c \mathcal{N}(\mu_{int,c}, \sigma_{int,c}) \mathcal{U}(0, n_{\text{bv}})}{p(int,x,y,z)} & \text{, when } c = 0, \\ \frac{\pi_c \mathcal{N}(\mu_{int,c}, \sigma_{int,c}) \mathcal{N}(\mu_{\mathbf{loc,c}}, \mathbf{\Sigma}_{\text{loc,c}})}{p(int,x,y,z)} & \text{, otherwise} \end{cases} \end{aligned} \tag{3.5}$$

For simplicity, we represent voxels using one-dimensional indices i. Here, $\gamma_{i,c}$ is the probability of voxel i belonging to cluster c, and π_c is the probability of cluster c. The cluster memberships are soft, i.e., each voxel can be a partial member of multiple clusters during the optimization process. We can then derive the update rules for the unknown parameters as follows (M-step):

$$\begin{aligned}
N_c &= \sum_{i=0}^{N} \gamma_{i,c} \\
\pi_c &= \frac{N_c}{N} \\
\boldsymbol{\mu}_{loc,c} &= \frac{1}{N_c} \sum_{i=0}^{N} \gamma_{i,c} \mathbf{coord}_i \\
\boldsymbol{\Sigma}^2_{\mathrm{loc,c}} &= \frac{1}{N_c} \sum_{i=0}^{N} \gamma_{i,c} (\mathbf{coord}_i - \boldsymbol{\mu}_{loc,c})(\mathbf{coord}_i - \boldsymbol{\mu}_{loc,c})^T \\
\mu_{int,c} &= \frac{1}{N_c} \sum_{i=0}^{N} \gamma_{i,c} int_i \\
\sigma^2_{\mathrm{int,c}} &= \frac{1}{N_c} \sum_{i=0}^{N} \gamma_{i,c} (int_i - \mu_{int,c})(int_i - \mu_{int,c})^T
\end{aligned} \tag{3.6}$$

where N_c is the effective number of elements in cluster c, π_c is the proportion of voxels belonging to cluster c, $\mathbf{coord_i}$ is a vector containing the coordinates for voxel i, and int_i is the intensity of voxel i.

Traditional EM algorithm iterates between the E- and M-steps until the solution converges. However, the algorithm expects that the number of clusters is known ahead of time, which is not the case with previously unseen CT scans. Therefore, we need to add an additional step to determine if there are any hemorrhage clusters in the beginning and if we should add more clusters to better represent the hemorrhage regions once the EM algorithm has converged.

First, we can use prior knowledge of the problem domain to determine which voxels we have a high certainty about. Typically, brain tissue has lower intensity values, so we start

by initializing all voxels with intensity values below 40 to belong in the healthy cluster. Determining which voxels contain hemorrhage is more challenging because high intensity values could be either hemorrhage or noise. Therefore, we look for large contiguous regions with high intensity values ($>$ 50 HU). If one exists, we create a new cluster containing these voxels. This allows us to calculate the initial cluster statistics and start optimizing the clusters by using the EM algorithm. After optimizing these clusters, we look for any remaining large high-intensity regions that do not belong to hemorrhage clusters, and we create new hemorrhage clusters for them. This process is repeated until we have no remaining high-intensity regions that do not belong in hemorrhage clusters. As a result of this approach, we can have a contiguous hemorrhage region that is represented by multiple clusters if the shape cannot be represented by a single Gaussian distribution (for example, crescent or V shapes). Therefore, the number of clusters does not necessarily correspond to the number of distinct hemorrhage regions.

3.3.4 Post-processing

After the EM algorithm has finished, we have a soft clustering where each voxel can belong in multiple different clusters with certain probabilities and there might be multiple hemorrhage clusters. Our goal, however, is to mark each voxel either as healthy or as hemorrhage. To do so, we take the sum of the probabilities of all hemorrhage clusters and compare it to the probability of the healthy cluster. If the combined probability of the hemorrhage clusters is higher than the probability of the healthy cluster, we mark the voxel as hemorrhage:

$$\text{label}_i = I(\sum_{c \geq 1} p(c|i) > p(c = 0|i)), \tag{3.7}$$

where I is the indicator function. Typically, post-processing is also needed to remove noise from the segmentation results. However, as the optimization process already takes into account the pixel neighbourhoods, i.e., a high-intensity voxel is only marked as hemorrhage if

it is located near other high-intensity voxels, our approach is unlikely to mark noisy voxels as hemorrhage. If a noisy voxel is located near hemorrhage, it would be difficult to distinguish from actual hemorrhage even for a human annotator. Therefore, we did not find it beneficial to perform noise removal with our algorithm. However, in some cases, the hemorrhage regions contain low-intensity holes, which we fill by performing morphological closing. This will fill the small holes while leaving the larger gaps intact, which corresponds with how a radiologist would outline the hemorrhage.

Our entire algorithm is outlined in Figure 3.3.

3.4 Experiments And Results

3.4.1 Quantitative Evaluation

We compare our algorithm to Fuzzy C-Means (FCM40, FCM45), PItcHPERFeCT [127], and DeepBleed [153]. FCM is an unsupervised algorithm that finds clusters based on the voxel intensity. It fits multiple Gaussian distributions on the intensity distribution and marks the clusters with a high mean intensity as hemorrhage. For the purposes of our comparisons, we provide results for FCM using two intensity thresholds: 40 and 45. The first threshold was chosen because it achieved the highest Dice scores in our experiments while the second threshold was chosen because it is less likely to include noise in the results. Our implementation of FCM uses the same preprocessing pipeline as our proposed algorithm and its results are postprocessed with morphological opening to eliminate noisy voxels. PItcHPERFeCT and DeepBleed are supervised models that have an open source implementation available, and we use the pretrained models published by the original authors.

We first evaluate these algorithms using the CQ500 dataset. The dataset contains 205 hemorrhage patients but we chose to exclude 9 of them who had chronic hemorrhage. Acute and chronic hemorrhages manifest differently and therefore need different approaches, and none of the algorithms in our comparison have been designed for detecting chronic hemor-

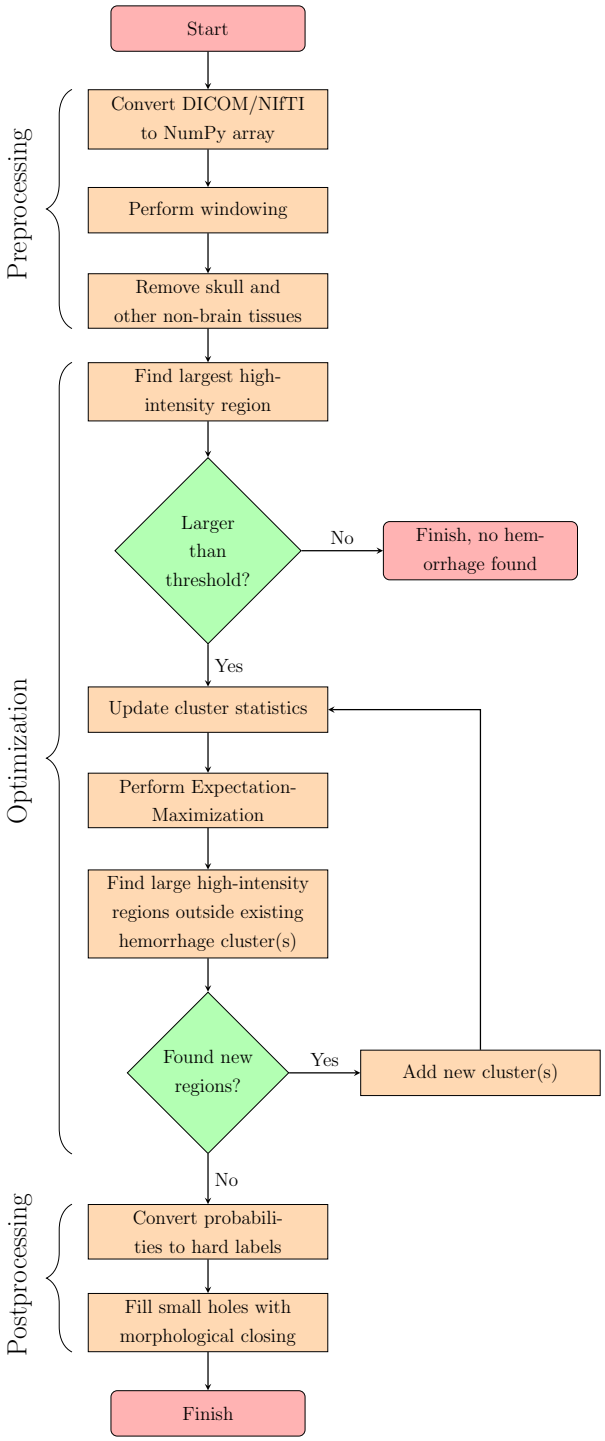


Figure 3.3: Hemorrhage segmentation process

rhages. As a result, we had 196 hemorrhage patients in our comparison for whom we had bounding boxes delineating the hemorrhage regions in each cross-sectional slice. To evaluate the correctness of the segmentation, we calculate the Dice score which determines the level of similarity between two segmentations and is defined as:

$$\text{Dice} = \frac{2 * |X \cap Y|}{|X| + |Y|}, \tag{3.8}$$

where X and Y are sets of voxels that have been marked as hemorrhage. While Dice score typically ranges from 0 (no matching voxels) to 1 (perfectly matching voxels), our experiments compare segmented images to bounding boxes which typically include healthy tissue as well. As a result, the score of a perfect segmentation would be lower than 1 in practice.

The results can be found in Table 3.1. We first show the results using all of the patients and then by dividing the patients into subgroups based on their hemorrhage types. As one patient can have one or more types of hemorrhage, we further divide the subgroups into patients who have a certain hemorrhage type and possibly other types as well, and into patients who have only a certain hemorrhage type.

Table 3.1: Dice scores on CQ500 dataset

	Our Model			FCM40			FCM45			PItcHPERFeCT			DeepBleed		
Hemorrhage type	Mean	Max	Std	Mean	Max	Std	Mean	Max	Std	Mean	Max	Std	Mean	Max	Std
All patients (N=196)	**.197**	**.814**	.222	.149	.563	.137	.110	.504	.129	.090	.556	.143	.141	.684	.184
Patients w/ intraparenchymal (N=137)	**.237**	**.814**	.218	.162	.476	.138	.128	.504	.136	.123	.556	.162	.207	.684	.200
Patients w/ subdural (N=64)	**.244**	**.733**	.241	.190	.563	.130	.139	.406	.119	.056	.345	.081	.066	.402	.101
Patients w/ epidural (N=6)	**.245**	**.499**	.180	.191	.364	.010	.176	.364	.128	.118	.342	.123	.142	.402	.145
Patients w/ subarachnoid (N=113)	**.221**	**.814**	.233	.175	.563	.140	.119	.432	.123	.074	.556	.113	.092	.563	.129
Patients w/ intraventricular (N=40)	**.363**	**.814**	.229	.272	.465	.129	.213	.504	.136	.170	.556	.175	.194	.594	.192
Patients w/ only intraparenchymal (N=45)	.128	.554	.172	.059	.386	.082	.045	.399	.096	.108	.507	.169	**.265**	**.684**	.226
Patients w/ only subdural (N=22)	**.197**	**.733**	.256	.177	.496	.127	.139	.406	.128	.040	.222	.063	.035	.189	.061
Patients w/ only epidural (N=1)	**.027**	**.027**	.000	.019	.019	.000	.032	.032	.000	.000	.000	.000	.000	.000	.000
Patients w/ only subarachnoid (N=28)	**.022**	**.275**	.065	.056	.244	.067	.017	.169	.041	.014	.163	.041	.012	.244	067
Patients w/ only intraventricular (N=0)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

As the results show, our model reaches higher mean and maximum scores than the other models in all but one hemorrhage type. In the case of intraparenchymal hemorrhage, DeepBleed is able to segment the hemorrhages more accurately than our model, but our model still outperforms FCM40, FCM45, and PItcHPERFeCT models. The most significant differences between our model and the supervised models occur with the subdural and subarachnoid hemorrhage types.

3.4.2 Detection Rate Analysis

To further understand the situations where the models provide differing results, we inspect the detection rate of bounding boxes based on the hemorrhage characteristics. First, we compare the results on different bounding box sizes (see Figure 3.4). The results show that all models perform better with larger hemorrhage sizes, which is as expected, as the smaller hemorrhages are more difficult to distinguish from noise. The size has a large effect especially on PItcHPERFeCT (increase from 4% to 56%) and DeepBleed (increase from 21% to 56%), while FCM models (86% to 99%, 48% to 88%) as well as our model (62% to 78%) are slightly less affected. While FCM40 provides the highest detection rate with all hemorrhage sizes, it should be noted that it is more likely to include noise as well, which is evidenced by the lower Dice scores.

When comparing the results on different maximum intensities (see Figure 3.5), we can again see that all the algorithms perform worse on low-intensity hemorrhages as expected. The most common situation where the intensity is low is when the hemorrhage is older, but in some less-frequent situations even acute hemorrhages can have a low intensity. The detection becomes more challenging in these cases, as the intensity is similar to healthy brain tissue's intensity, and therefore a radiologist might have to look for other signs of hemorrhage. As FCM40 and FCM45 are focused on the voxel intensities, their detection rate increases rapidly as the intensity level increases. Our model's detection rate increases very early as well, while DeepBleed and PItcHPERFeCT reach high detection rates only when the voxel

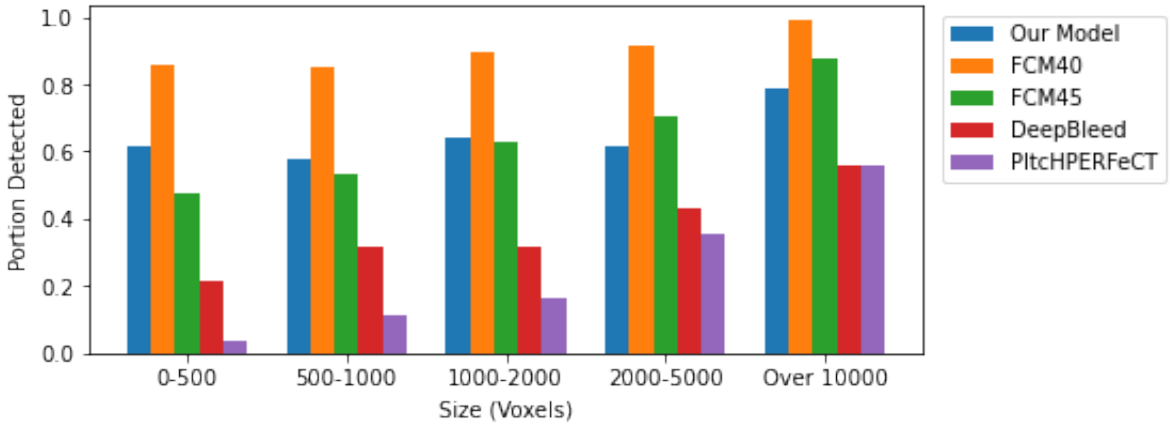


Figure 3.4: Comparison of hemorrhage detection rates based on bounding box size.

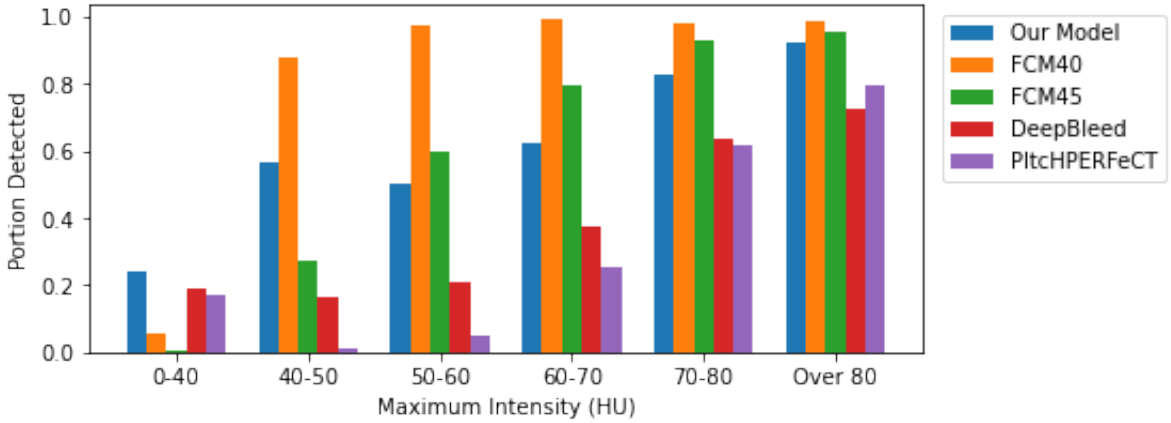


Figure 3.5: Comparison of hemorrhage detection rates based on maximum intensity.

intensities are over 70 HU.

Finally, when looking at the hemorrhage types (see Figure 3.6), we can see that the FCM models provide similar results on each type. This is to be expected, as they only look at voxel intensities without taking the shape or location of the hemorrhage into account. PItcHPERFeCT and DeepBleed show lower detection rates on subarachnoid and subdural hemorrhages, which matches with our observations on the lower Dice scores as well. Our model shows small variations on the detection rates with the highest detection rates occurring on subarachnoid and epidural hemorrhages.

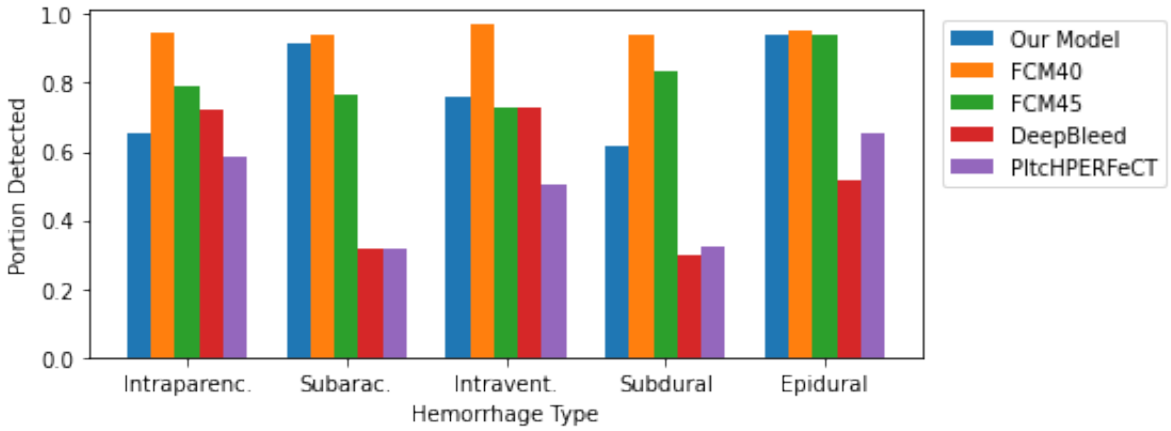
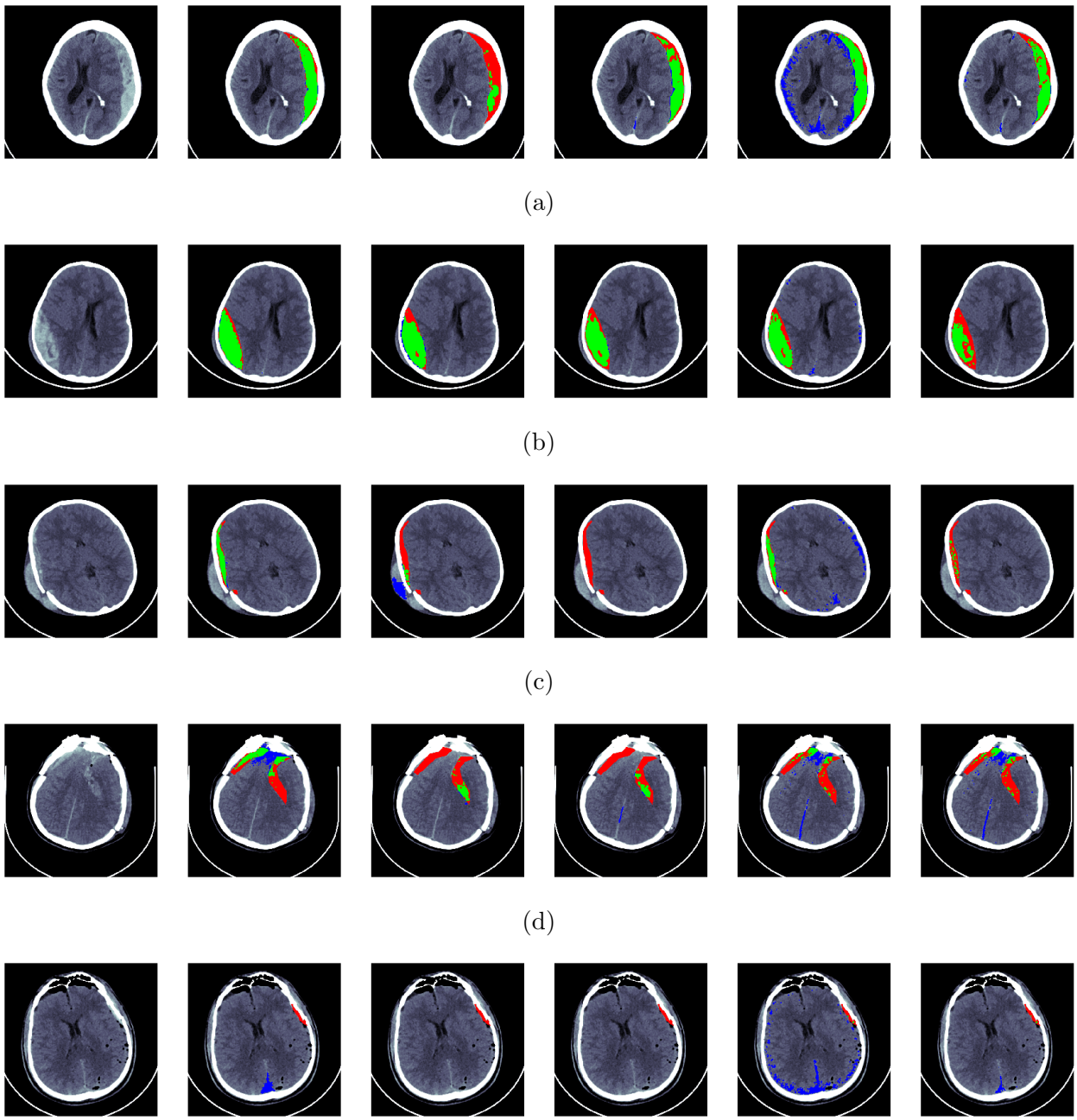


Figure 3.6: Comparison of hemorrhage detection rates based on hemorrhage type.

3.4.3 Visual Evaluation

Next, we perform visual analysis on the results to better understand where each model succeeds or fails. For the visualizations, we use the smaller PhysioNet dataset which provides ground truth segmentations so we can see precisely how each model's results differ from a radiologist's segmentation. We have hand-selected a set of CT scans that contains both successful and failed segmentations. The original and segmented images can be seen in Figure 3.7. It should be noted that all of the chosen algorithms perform their detection on the full three-dimensional CT scan even though we only show one slice from each scan.

First, Image 3.7a shows a wide subdural hemorrhage region on the right side along the skull. Our model matches the ground truth very accurately and only misses a small region at the top part of the hemorrhage as well as narrow slices along the edges. In the top part of the image, the hemorrhage's intensity becomes very similar to other brain tissue's intensity which makes it challenging to detect correctly. DeepBleed fails to detect most of the hemorrhage while PItcHPERFeCT only detects the highest-intensity regions of the hemorrhage leaving gaps at the locations that have a lower intensity. FCM40 detects most of the hemorrhage but includes large noisy regions along the edges of the brain and FCM45 misses the low-intensity regions of the hemorrhage.



(a)

(b)

(c)

(d)

(e)

Figure 3.7: Visual evaluation of the models. From left to right: 1) Original image, 2) Our model, 3) DeepBleed, 4) PItcHPERfeCT, 5) FCM 40, 6) FCM 45. Green indicates correctly segmented voxels, red indicates false negative voxels, and blue indicates false positive voxels.

Image 3.7b contains epidural hemorrhage on the lower-left part of the image. Again, our model matches the ground truth nearly perfectly while only missing a narrow slice along the right edge of the hemorrhage. DeepBleed and PItcHPERFeCT both leave out the top part of the hemorrhage, most likely due to the slightly lower intensity within that area. FCM40 misses the areas along the top-right edge of the hemorrhage while marking small regions along the edges of the brain as hemorrhage. FCM45 does not detect any incorrect regions but misses large parts of the hemorrhage due to the lower intensity.

Image 3.7c shows a subdural hemorrhage close to a skull fracture. Our algorithm detects most of the hemorrhage but misses a small, detached region below the fractured area. DeepBleed and PItcHPERFeCT miss most of the hemorrhage presumably because it is a very narrow region and has only a slightly higher intensity than the healthy brain tissue. DeepBleed also marks part of the edema located outside the skull as hemorrhage which indicates that the brain extraction did not work correctly with a fractured skull. Again, FCM40 misses part of the hemorrhage while including noise and FCM45 misses most of the hemorrhage.

Image 3.7d demonstrates a more complicated situation where the skull is fractured and significant parts of the hemorrhage have a somewhat similar intensity to the other brain tissue. As the top-most part of the hemorrhage is close to higher-intensity brain tissue, our model mistakenly marks some of the healthy tissue as hemorrhage. At the same time, our model does not detect the lower regions of the hemorrhage which have a lower intensity. DeepBleed and PItcHPERFeCT show the opposite behavior, as they only detect the lower region while missing the top parts of the hemorrhage. Both FCM40 and FCM45 perform poorly because the hemorrhage's intensity is very close to other brain tissue's intensity. Note especially the midline which has a higher intensity than most of the actual hemorrhage and which was incorrectly detected by FCM40, FCM45, and PItcHPERFeCT

Finally, Image 3.7e shows a situation where all the models fail. It contains a very narrow hemorrhage region along the right side of the skull, which is hard to distinguish from noise,

as even healthy patients typically have a slightly higher intensity near the edges of the brain. Instead of detecting this region, our model incorrectly detects the lower region which has a high intensity. DeepBleed and PItcHPERFeCT do not detect anything in this image while FCM40 and FCM45 incorrectly mark other regions near the edges as hemorrhage.

3.5 Discussion

As was shown in the previous section, our model can achieve similar or better results than the existing models on most hemorrhage types. Supervised models are typically expected to outperform unsupervised models due to their capability of learning more complicated patterns, but to our surprise, there was only one hemorrhage type where DeepBleed performed better than our model. Even in that category, our model outperformed the remaining models. There are multiple possible explanations for why the supervised models did not perform better. For example, their training data could differ from the public datasets that we used for evaluating the models, but it cannot be verified, as the training data has not been released. These possible differences could include, for example, a different distribution of hemorrhage types, different kinds of images (e.g., more high-intensity hemorrhages), or different CT scanner models. These differences could lead to the models overfitting to certain types of images and not generalizing well to other datasets.

When comparing the bounding box detection rates, the supervised models were significantly worse when the size was small. Our model's results could be explained by the fact that many of the smaller bounding boxes tend to be near the edges of a larger hemorrhage region. Due to our formulation, these regions are easily distinguished from noise as they are detected as a part of the larger region. This same can also explain part of the differences in the detection rates of the low-intensity bounding boxes.

Our model was also able to perform better than the FCM models. This was the expected outcome, as the FCM models are focused on the intensity levels, while our model is able to

take into account the voxel neighborhood as well. As can be seen from the visual comparisons in the previous section, our approach was able to avoid most of the high-intensity noise while simultaneously including the low-intensity parts of the hemorrhage as long as it was connected to a larger hemorrhage region. In general, our algorithm shows a good ability to adapt to hemorrhages with different intensities and shapes due to how the hemorrhage distributions are represented in our model.

Some of the most significant remaining challenges are 1) very small hemorrhage regions, 2) nearly-isodense hemorrhages, and 3) high-intensity healthy regions. The very small hemorrhages are not detected due to our algorithm's minimum threshold for hemorrhage sizes, which is necessary to avoid including noise. Detecting some of the smaller hemorrhages could be possible by requiring a larger intensity if the hemorrhage is small, but even this approach would be missing some hemorrhages which are small and have a low intensity. Detecting nearly-isodense hemorrhages and ignoring high-intensity healthy regions are closely related problems. In both cases, hemorrhage and healthy tissue look very similar, and correct detection might require additional information. For example, it could be possible to take into account that certain regions of the image are likely to have higher intensities in general, for example, the edges and the midline. The algorithm could have different intensity requirements for these regions, but it requires further analysis of the image to detect, for example, the midline correctly. For simplicity, we have not included this in the proposed algorithm.

While we have shown that unsupervised techniques can perform well in many situations, to further improve performance, future work should explore semi-supervised techniques that can leverage a small number of labeled images together with a larger number of unlabeled images. Hospitals may have access to a vast quantity of prior medical images, but manually labeling all of them is not practical. Semi-supervised techniques can help utilize the unlabeled images with a limited number of labeled images. Additionally, by obtaining data sharing consent from a small group of patients, multiple hospitals could use the same publicly available labeled images together with their private data to train highly accurate models

without compromising patient privacy. This approach may be more feasible than sharing entire training datasets.

3.6 Conclusion

In this chapter, we have presented a novel unsupervised algorithm for segmenting intracranial hemorrhage within CT scans. This allows radiologists to perform their work more efficiently because there can be dozens of images per CT scan, each of which has to be inspected separately. While many supervised techniques have been proposed for this problem, the lack of publicly-available training data makes them inaccessible to most. Our proposed algorithm is based on the mixture models and can adaptively choose the appropriate number of clusters so that the hemorrhage is represented accurately. We have provided a comparison of our algorithm and a number of earlier algorithms which shows that our results are consistently better than the prior algorithms in nearly all of the categories.

CHAPTER 4

Sleep and Activity Prediction for Type 2 Diabetes Management Using Continuous Glucose Monitoring

4.1 Introduction

As of 2021, an estimated 536.6 million people worldwide had diabetes, and the number is expected to increase [166]. Among people with diabetes, as many as 90-95% have Type 2 diabetes (T2D) [35], which appears to be caused by complex interactions of genetic, environmental, and lifestyle factors, such as lack of physical activity and being overweight. T2D is characterized by insulin resistance and pancreatic beta-cell dysfunction which leads to higher blood glucose levels once the pancreas can no longer keep up with increased demand [131]. While medical care is improving, people with T2D are at high risk for serious microvascular (diabetic retinopathy, neuropathy, nephropathy) and macrovascular (cardiovascular disease, peripheral artery disease) complications [36]. Costs incurred due to diabetes-related medical expenses and reduced productivity were estimated to be $327 billion in the US in 2017 [7].

Daily actions, such as physical exercise, healthy eating, and sufficient sleep, are important factors in both preventing and delaying the onset of T2D as well as its management. For example, The Diabetes Prevention Program [171] was able to reduce diabetes incidence rates by 57% with a lifestyle intervention which included weight reduction/maintenance and regular physical activity. The American Diabetes Association's (ADA) Standards of Care (SOC) recommends at least 150 minutes of moderate-to-vigorous physical activity per week spread out evenly throughout the week [8]. The relationship of sleep with diabetes has

received little attention in the continuous glucose monitoring (CGM) literature, but there is a clear bi-directional relationship between sleep and T2D. Multiple studies have shown that sleep restriction, poor sleep quality, and irregular sleep cycles can lower insulin sensitivity [148,158] while increasing sleep quantity has been shown to improve insulin sensitivity among healthy individuals [106]. Moreover, poor sleep habits and sleep disorders are highly prevalent among adults with T2D, and low-quality sleep is associated with poor diabetic outcomes. Conversely, T2D and its common complications may adversely affect sleep and sleep quality.

In recent years, CGM data has been used in many ways to assist people with diabetes, with the main focus being on Type 1 diabetes (T1D). Much of this work has focused on preventing dangerous conditions such as hypoglycemia [16,55,94,118,150] and hyperglycemia [118], in which the blood glucose level goes outside the safe range. These techniques can provide individuals with an early warning or they can directly control their insulin pump to prevent potentially dangerous scenarios before they occur. Some studies have also predicted the glucose levels directly [80, 177]. In addition, CGM data has been used to determine additional insulin dose requirement at mealtimes [132].

In this chapter, we demonstrate the strong connection between CGM signal and daily activities, as inferred by an activity tracker. We do this by training a neural network model which can predict sleep, walking, and elevated heart rate directly from the CGM signal. Being able to predict activities may expand the utility of the CGM device in diabetes management. In addition, this can be used to analyze CGM and activity jointly to better understand their relationship and effects on health outcomes. We then develop the model further to include demographic information as well as medical claims data, and we show that this additional information further improves the physical activity prediction results.

A few prior studies have used CGM data to determine people's daily actions. For example, CGM data has been used to determine adherence to daily insulin injections [172], and to detect eating activities [17, 136]. We are not aware of any studies that have predicted sleep from CGM data and to the best of our knowledge, only one study has used CGM data to

predict physical activity [11]. In that study, self-reported meal and exercise activities were predicted for 11 people with T1D using their glucose and insulin data. A separate recurrent neural network model was trained for each person using a part of their data and the model was evaluated by predicting activities in the remaining data. This differs from our study in multiple ways: our focus is on T2D, we demonstrate the results using orders of magnitude larger dataset, and we do not assume that any training data is available for an individual before making predictions, thus making our results more generalizable.

The main contributions of this chapter are:

1. Demonstrating on a cohort of 6981 people with T2D that the sequence of daily activities, such as sleep and physical activity, can be determined from a CGM signal without the need for additional activity tracking devices.

2. Designing a neural network architecture for time-series data which can also utilize static or low-frequency variables, such as demographics and medical claims data, and demonstrating that this additional data improves the physical activity predictions compared to a model which only uses CGM data.

4.2 Methods

4.2.1 Dataset

Our dataset consisted of de-identified CGM, fitness tracker, medical claims, and demographic data collected from individuals in a commercial diabetes care program between October 2019 and April 2022. Individual medical histories were captured by medical claims starting from January 2016 to April 2022. Individuals in this program had T2D diagnosis in the last 24 months or had been prescribed certain glucose-lowering medications, and the program provided them lifestyle interventions, digital self-care tools, and wearable technology with the goal of increasing glucose time in range, lowering HbA1c, and minimizing reliance on

medication where possible. As some individuals did not wear the devices consistently, we ensured data quality by only using data for calendar days where the individual had at least 22 hours of overlapping CGM and fitness tracker data. In addition, we required them to have at least 1 hour of sleep during that day to ensure the fitness tracker was collecting data accurately. We used data for all individuals who had at least one day of data.

4.2.2 Data Preprocessing

Glucose Data

Glucose level was measured using a CGM device which provided interstitial glucose measurements every 5 minutes. In addition to glucose measurements, the device calculated the current trend, i.e., how rapidly the glucose value was increasing or decreasing. We only used days that had at least 22 hours of overlapping glucose and fitness tracker data, and missing values were imputed using linear interpolation.

Activity Data

Physical activity and sleep data were collected using a Fitbit fitness tracker which reported the average heart rate and step count once a minute, as well as the intervals when the individual was determined to be asleep based on their heart rate and accelerometer data. Systematic reviews have shown that Fitbit devices similar to ones used in our study are fairly accurate at detecting sleep and measuring heart rate and steps. However, they may slightly overestimate the total sleep time (9-11.6 minutes on average) [77], and they may slightly underestimate heart rate (-2.99 beats per minute) and steps (-3.11 steps per minute) [45].

Due to the difference in time sampling frequencies compared to the CGM data, fitness tracker data was aggregated by taking the average measurements for each 5-minute time block to match the CGM. Sleep was treated as a boolean value which indicated whether the individual was asleep or awake at the beginning of each 5-minute block. Missing activity

labels were also imputed using linear interpolation.

Heart rate values were converted to heart rate zones because the absolute heart rate values corresponding to different activity levels can differ significantly between individuals. This conversion was performed using personalized zones determined by the fitness tracker, which determined them by using an estimate of the maximum heart rate as well as the measured resting heart rate. The zones were Out of Range (lowest), Fat burn, Cardio, and Peak (highest). These zones were further binarized to two values: Fat burn or higher, and Cardio or higher. These two zones were approximately equivalent to moderate and vigorous activity levels that are used in physical activity guidelines. We did not create a separate class for Peak zone because of the very small number of individuals reaching high enough heart rates (on average, 2% of individuals' days had Peak activity).

Walking activity was binarized using a threshold of 400 steps per five minutes. Typical moderate-intensity walking pace is about 100 steps per minute [155], so we chose a limit that was slightly below it to reliably capture all moderate-intensity walking even if the individual had to stop for a very short time (e.g., waiting to cross a road).

To reduce noise in the activity labels, we filtered out individual positive values, thereby only considering activities that lasted 10 minutes or longer at a time. This time limit was chosen to be the same as the American Diabetes Association's recommendation for the minimum duration of a physical activity [8].

Medical Claims Data

We accounted for medical history by means of medical claims data. Medical claims were de-identified administrative claims data for Medicare Advantage and commercially insured individuals. This database contained medical (emergency, inpatient, outpatient) claims for services submitted for third party reimbursement, available as International Classification of Diseases (ICD) [184] claims. These claims were aggregated after completion of care encoun-

ters and submission of claims for reimbursement. We grouped similar ICD codes together by only considering the first 3 letters/numbers of each code.

Demographics Data

Demographics data included each individual's age, gender, height, weight, and BMI. Age and gender were collected at the time of enrollment and were assumed to remain constant (very few individuals had more than a year of data), while height and weight were self-reported using the fitness tracker app and the values might have changed over the data collection period. If the individual did not report any height or weight measurements, we imputed them using the mean value. BMI was calculated using the height and weight measurements.

4.2.3 Model

Our model was based on the U-Net architecture (see Figure 4.1) [146], which was designed for 2-dimensional medical image segmentation. Modified versions of the U-net architecture have been shown to be effective for a wide variety of tasks, such as volumetric medical image segmentation [53, 122], image denoising [93, 112, 113, 137], audio source separation [91,159,169], speech enhancement [49,139], and physiological signal imputation [23,30,38,82].

As our goal was to transform a sequence of glucose measurements to a sequence of daily activities, we modified the original U-net architecture to use 1-dimensional convolutions instead of 2-dimensional convolutions. The input matrix $\mathbf{X}$ was a sequence of 288 glucose measurements (24 hours * 12 measurements/hour) along with the rate of change computed by the CGM device. These two values were combined as separate channels, i.e., $\mathbf{X} \in \mathbb{R}^{2\times 288}$. Prediction targets included four activities: sleep, walk, heart rate in fat burn zone or higher, and heart rate in cardio zone or higher. These values were measured at the same time intervals as the glucose values and treated as separate channels, therefore the target matrix $\mathbf{Y} \in \mathbb{R}^{4\times 288}$. The prediction was treated as a multi-label classification task, i.e., multiple

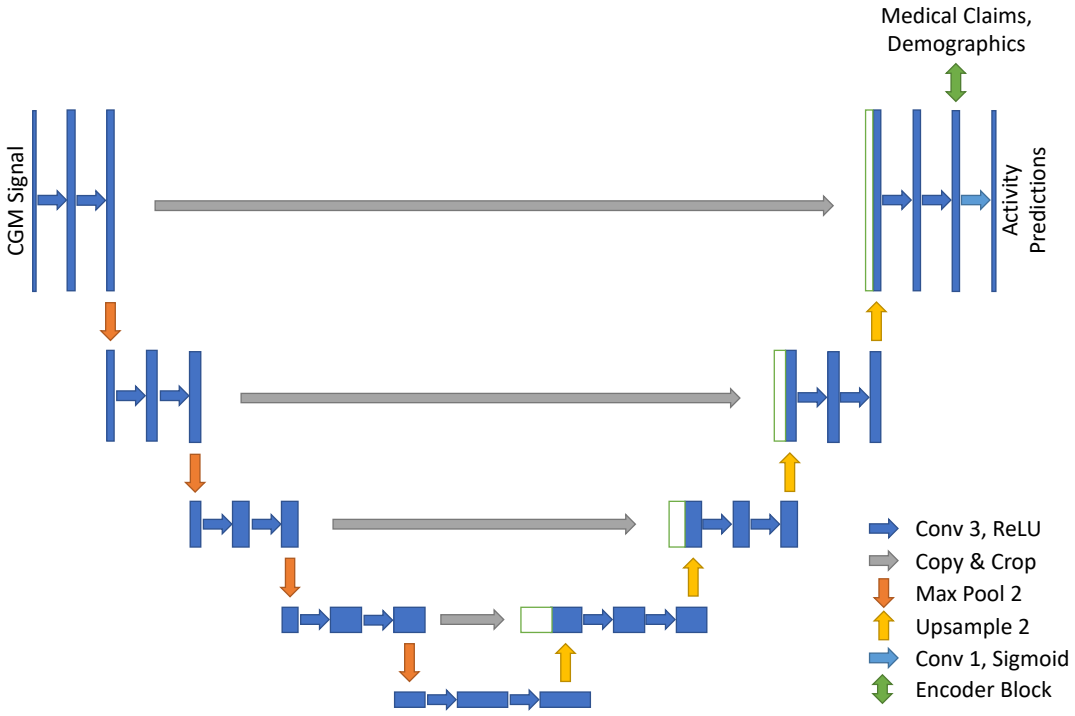


Figure 4.1: Modified U-Net architecture with an encoder for demographics and medical claims.

target values were allowed to be positive simultaneously (e.g. walking and high heart rate).

Next, to incorporate claims data into the model, we converted the ICD codes that occurred prior to the current time window into vectors $\mathbf{x_{icd}} \in \mathbb{R}^{300}$. This conversion was done using pretrained ICD2Vec embeddings [105], which transformed the high-dimensional one-hot encoded data into a lower dimensional mapping. These embeddings were trained using text descriptions of ICD codes and therefore ICD codes with similar descriptions had similar representations. As our hypothesis was that some ICD codes might be more relevant for the prediction task than others, we wanted the model to be able to learn which ICD codes to use. In addition, we wanted the model to be able to distinguish between acute and chronic diseases, as acute diseases might be relevant only for a few days or weeks while chronic ones could have effects across many years. Including ICD codes into the model was implemented using a multi-head attention layer [178] as shown in Figure 4.2. Time was first converted

into larger time blocks to avoid overly sparse time values (see Figure 4.3). These time blocks were generated using conversion:

$$f(t) = \lfloor \log(t) \rfloor \tag{4.1}$$

where t is the number of days between the current time window and the claim, and $\lfloor \cdot \rfloor$ is the floor function. This time block was then converted into a vector $\mathbf{t} \in \mathbb{R}^{50}$ using step encoding shown in [178]:

$$\begin{aligned} \mathbf{t}_{\text{pos},2i} &= \sin\left(\frac{\text{pos}}{10000^{2i/d_{\mathbf{t}}}}\right) \\ \mathbf{t}_{\text{pos},2i+1} &= \cos\left(\frac{\text{pos}}{10000^{2i/d_{\mathbf{t}}}}\right) \end{aligned} \tag{4.2}$$

where $d_{\mathbf{t}}$ is the dimensionality of vector $\mathbf{t}$. The resulting vector was concatenated to the ICD vector $\mathbf{x_{icd}}$. The resulting vector was used as both the key $\mathbf{k}$ and value $\mathbf{v}$ for the attention layer while a constant vector was used as the query $\mathbf{q}$.

We evaluated including the combined ICD representations in three locations within the U-Net model: in the input layer, between encoder and decoder, as well as before the last convolutional layer. We call these locations early, middle, and late in the results section. These vectors were concatenated with the U-Net's internal representation and passed through a convolution with size 1 and a rectified linear unit (ReLU) layer. The resulting vector was then returned to the same location of the U-Net model where the original vector was. Demographics (age, gender, height, weight, BMI) were standardized and combined with the internal representations the same way.

4.2.4 Training

We randomly divided individuals 60%-20%-20% to training, validation, and test sets respectively. Models were trained using AdamW [116] algorithm with learning rate 1e-3 for 35

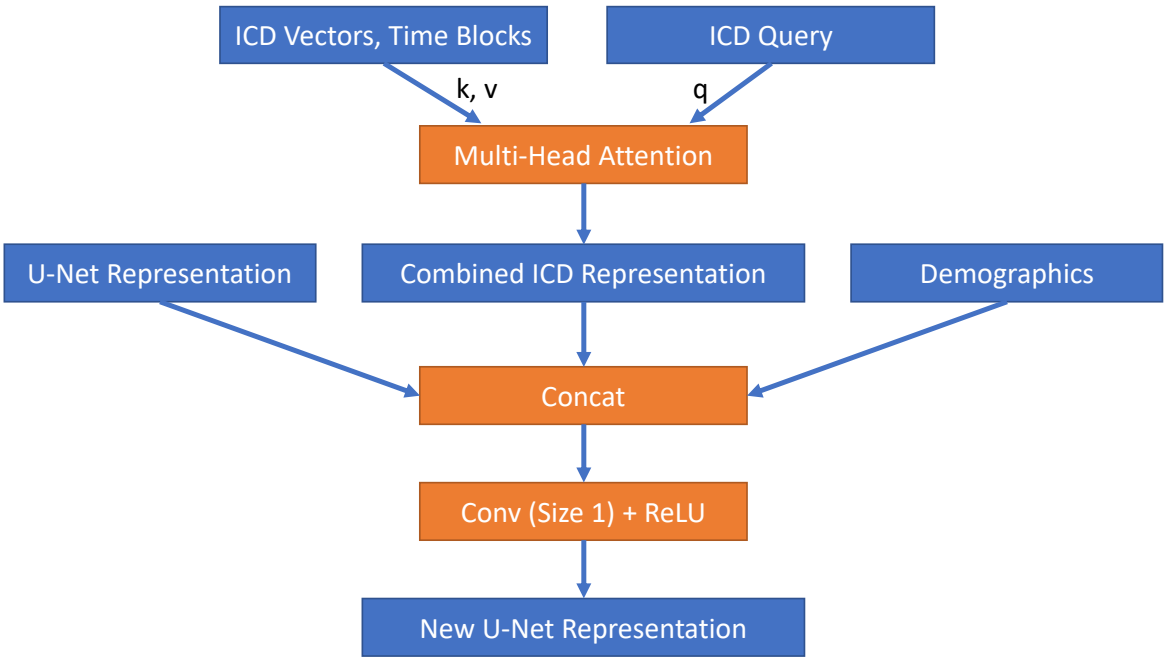


Figure 4.2: Combining ICD vectors using Multi-Head Attention and adding the combined vector along with demographics to U-Net's internal representation. The results were returned to the U-Net model at the same location where the original internal representation was.

epochs, which was enough to reach convergence. In each training epoch, a random day of CGM data was chosen for each individual such that each individual appeared exactly once in each epoch but different epochs might have used different time windows. In validation and test phases, one day of data was picked for each individual to be used in every epoch and experiment to ensure that the results were consistent across different experiments and each individual had an equal impact on the results.

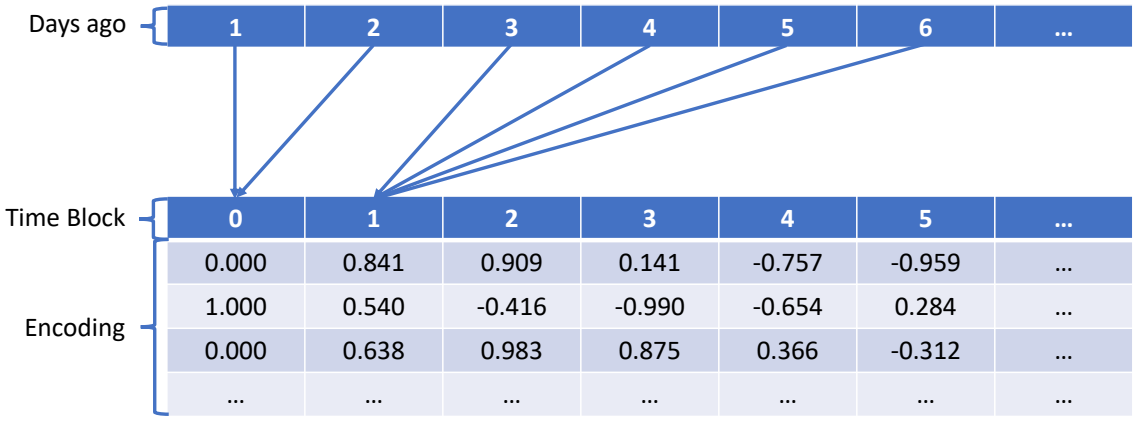


Figure 4.3: Claim dates were incorporated into the model by considering how many days prior to the current time window the claim occurred. The numbers of days were grouped into larger time blocks, which were then converted into time encodings.

4.3 Results

4.3.1 Dataset Statistics

Our dataset consisted of 6981 individuals, with an average of 65.1 days of data (median 28, standard deviation 85.6). The mean age of individuals was 54.3 years (median 55, SD 9.1), distribution of female/male was 51.3%/48.7%, and the mean body mass index (BMI) was 34.7 (median 33.4, SD 8.9). On average, individuals had medical claims on 64.5 unique days (median 49, SD 56.8), and a total of 172.9 ICD codes (median 127, SD 166.4), 40.9 of them being unique (median 36.0, SD 24.1).

4.3.2 Model Evaluation

We trained the model to predict a sequence of daily activities (sleeping, walking, heart rate in fat burn zone or higher, heart rate in cardio zone or higher) from a sequence of CGM values. We first evaluated our model with and without using demographics and ICD codes. We also performed an ablation experiment to determine how the results changed based on

the location where demographics and ICD codes were incorporated. We used one day of data for each person in the test set and bootstrapped the test data 100 times to determine the mean Area Under Receiver Operating Characteristics Curve (AUROC) along with the confidence intervals. Results of these experiments can be seen in Table 4.1.

As the results show, CGM data was highly predictive of sleep, and incorporating demographics or ICD codes did not provide significant improvements. In addition, CGM data was moderately predictive of walking and increased heart rate. Incorporating ICD codes and demographics was beneficial to physical activity predictions, and incorporating them in the later parts of the model was found to be better than incorporating them in the early parts. We hypothesize that incorporating ICD codes and demographics improved physical activity predictions because they can provide information about overall activity levels (e.g., a health condition which prevents physical activity) and how heart rate is affected by physical activity (e.g., due to age and health conditions). The lower results for heart rate predictions may be caused by the noisy labels, as our heart rate zones are based on personalized estimates.

Due to the significant class imbalance, we also evaluated area under precision-recall curve (AUPRC) scores, which show a similar pattern. The highest AUPRC score for sleep prediction was 0.884 (2.9 times higher than prevalence), for fat burn zone prediction 0.401 (1.9 times higher than prevalence), for cardio zone prediction 0.026 (3.0 times higher than prevalence), and for walk prediction 0.052 (8.7 times higher than prevalence).

Figure 4.4 shows an example of an individual's glucose values and activity predictions for one day. This figure shows that the model was very confident in the sleep predictions except in the beginning and end of the sleep period. The lower confidence in both ends can be partially explained by the noise in the ground truth data, as the sleep times were estimated by the fitness tracker based on movement and heart rate patterns. Walking probabilities were skewed toward zero because of the rarity of the event, but despite the miscalibration the model did give higher probabilities when the individual was truly walking. However, the probability sometimes increased at other times also, possibly when the individual was doing

Table 4.1: Comparison of AUROC scores using demographics and ICD codes at different parts of the model. The first row shows results without demographics or ICD codes, the following rows show results when either ICD codes or demographics or both are added at different locations (early, middle, late).

Demographics	ICD Codes	Sleep AUROC	HR Fat Burn AUROC	HR Cardio AUROC	Walk AUROC
-	-	0.946 (±0.001)	0.708 (±0.001)	0.727 (±0.004)	0.776 (±0.003)
-	Early	0.936 (±0.001)	0.687 (±0.001)	0.707 (±0.003)	0.770 (±0.004)
-	Middle	0.945 (±0.001)	0.704 (±0.001)	0.715 (±0.004)	0.790 (±0.003)
-	Late	0.945 (±0.001)	0.719 (±0.001)	0.744 (±0.004)	0.804 (±0.003)
Early	-	0.942 (±0.001)	0.669 (±0.001)	0.740 (±0.004)	0.809 (±0.002)
Middle	-	0.945 (±0.001)	0.716 (±0.002)	0.758 (±0.006)	0.797 (±0.003)
Late	-	0.946 (±0.001)	0.716 (±0.001)	0.752 (±0.004)	**0.820** (±0.003)
Late	Late	**0.947** (±0.001)	**0.722** (±0.001)	**0.768** (±0.005)	0.817 (±0.003)

other physical activities which had a similar impact on the glucose measurements as walking. Similarly, heart rate predictions had higher values when the heart rate was elevated, but the predictions were slightly noisy.

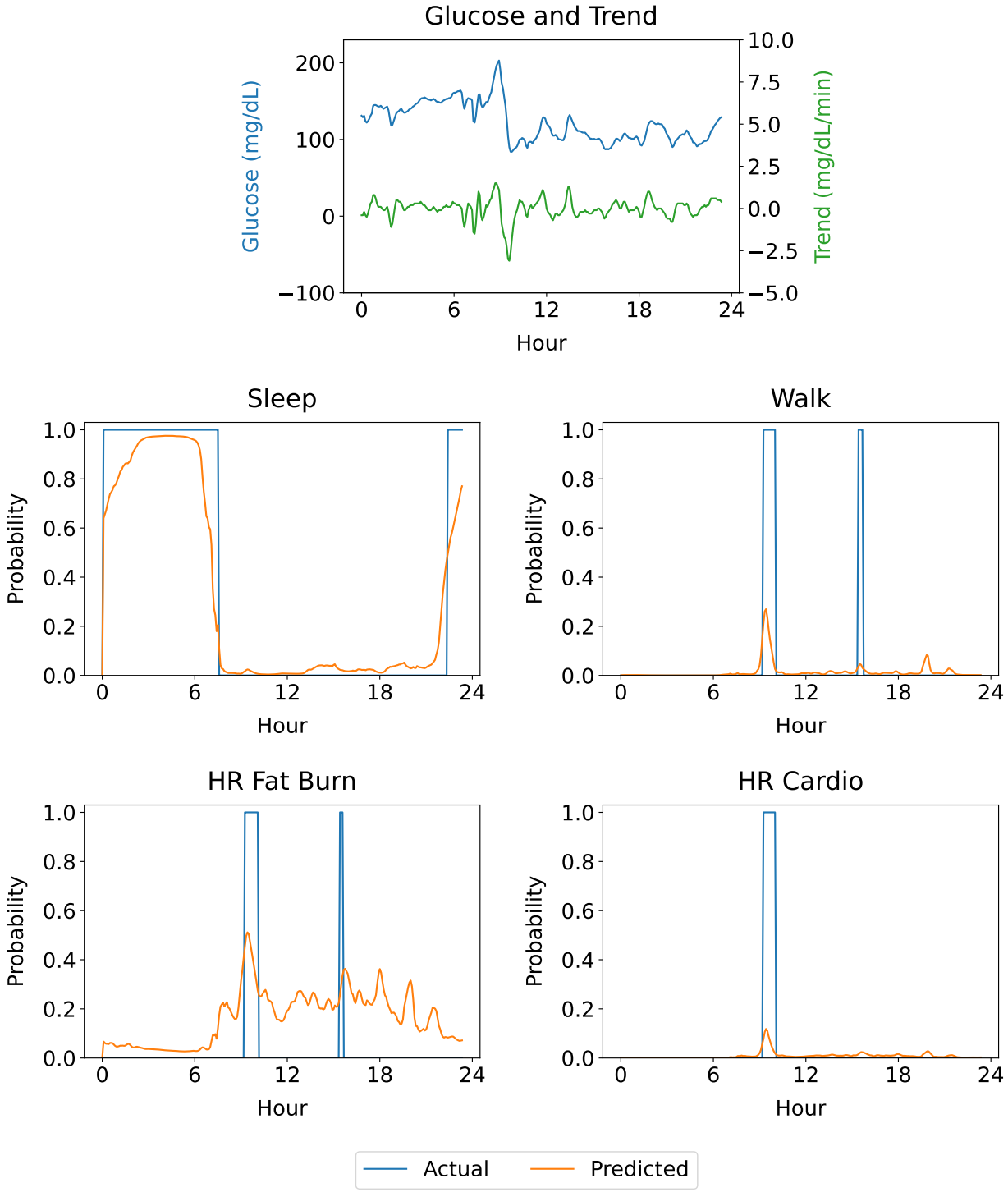


Figure 4.4: One individual's CGM signal and predictions for one day. X-axis shows hours starting from midnight.

4.3.3 ICD Attention Weights

Next, we analyzed how the best-performing model was using ICD codes. ICD codes were incorporated through an attention layer which learned how much weight to give to each ICD embedding while also taking into account how long ago the ICD codes appeared. To determine which ICD codes the model was relying on the most, we provided all ICD codes to the attention layer simultaneously at all time points and compared their relative attention weights. The highest attention weights and their corresponding ICD categories are shown in Table 4.2.

We found that the model was giving the most attention to various infections that had occurred recently as well as leg injuries which had occurred a long time ago. It should be noted that while the table shows only one time block per ICD code, the weights changed smoothly over time and the neighboring time blocks typically had very high attention weights also. Therefore, for example, more recent leg injuries had a significant impact on the predictions as well. We hypothesize that the recent infections were relevant because they increased the heart rates, and they might have also reduced the amount of physical activity. Infections can also cause metabolic changes [165] which could have altered the CGM data. Leg injuries likely reduced the amount of walking and other physical activities. Interestingly, the model also gave relatively high attention to code S69 (Wrist, Hand and Finger Injuries), which could be explained by different types of injuries having very similar pretrained embeddings, as the pretraining relied on the text descriptions of the codes. Alternatively, wrist injuries could have also affected the reliability of the data collected by the fitness watch.

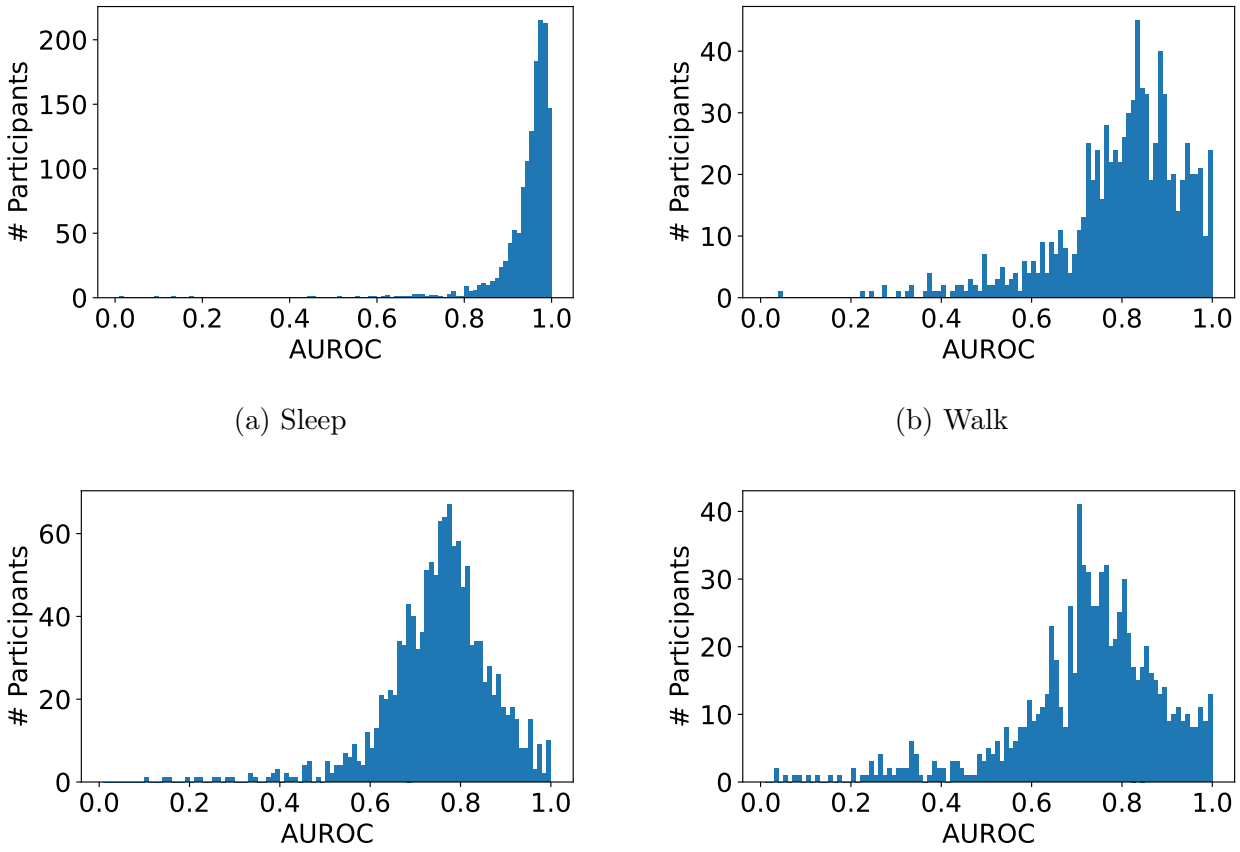
4.3.4 Differences Between Individuals

To determine how the model performed on individual people, we computed AUROC scores for each individual in the test set using their entire data. The distribution of individuals' AUROC scores is shown in Figure 4.5. As can be seen, sleep predictions were very accurate

Table 4.2: Highest relative attention weights given to different ICD categories. Time Block column shows which time block received the highest attention weight for this ICD code.

ICD-10 Code	Time Block	Attention Weight ($\times 0.01$)
J04 – Acute Laryngitis and Tracheitis	1 (3-7 days)	0.82
J22 – Unspecified Acute Lower Respiratory Infection	1 (3-7 days)	0.61
J02 – Acute Pharyngitis	1 (3-7 days)	0.43
J01 – Acute Sinusitis	1 (3-7 days)	0.41
S99 – Other and Unspecified Injuries of Ankle and Foot	7 (Over 1097 days)	0.41
J06 – Acute Upper Resp Infections of Multiple and Unsp. Sites	1 (3-7 days)	0.38
H65 – Nonsuppurative Otitis Media	1 (3-7 days)	0.27
S89 – Other and Unspecified Injuries of Lower Leg	7 (Over 1097 days)	0.26
S80 – Superficial Injury of Knee and Lower Leg	7 (Over 1097 days)	0.25
S69 – Other and Unspecified Injuries of Wrist, Hand and Finger(s)	7 (Over 1097 days)	0.25

for most of the people, with 89.6% having AUROC scores of 0.9 or higher. Walk predictions were also good for many individuals, with 26.4% having an AUROC score of 0.9 or higher and 62.4% having a score of 0.8 or higher. It should be noted that these distributions did not take into account individuals who never had positive values (i.e., sedentary individuals who were never physically active for 10 minutes at a time), as AUROC score is not defined in that situation. In those cases, it would have always been possible to choose a high enough threshold to get perfect (negative) predictions, so the distributions might underestimate the real accuracy of the models to some extent.



(a) Sleep

(b) Walk

(c) HR Fat burn zone

(d) HR Cardio zone

Figure 4.5: Distribution of individuals' AUROC scores on each task.

We also evaluated which demographic factors differentiated people who had high AUROC

scores (highest 25%) from people who had low scores (lowest 25%). In every task, the model made slightly better predictions for women, although the only statistically significant differences were in sleep prediction (highest AUROC scores contained 54% women, compared to 45% in the lowest 25% group, p=0.025) and fat burn zone prediction (56% women vs. 46% women, p=0.009). In addition, the group with high sleep prediction scores contained slightly younger people (52.6 years vs. 54.9 years, p=0.001) and slightly lighter people (224lbs vs. 233lbs, p=0.03).

4.3.5 Ablation Study

To determine which factors affected our prediction results, we evaluated various modifications of our model. Results for these experiments are shown in Table 4.3.

Our first experiment evaluated whether a more simple model could reach good prediction results as well. We tested this by training a neural network with a single linear layer which was trained the same way as our proposed model. As the results show ("Linear model" in the results table), our model performs significantly better than a simple model. However, a simple model was still able to make decent predictions in some cases, especially when predicting sleep. We hypothesize that these predictions rely heavily on the time of day rather than the CGM signal itself, as the linear model is not capable of capturing more complicated patterns in the signal.

To evaluate how much our model depended on the time of day, we performed two experiments. In the first experiment ("Randomized time"), we randomized the time windows by shifting each individual's data by 0-23 hours at random. A new model was trained using this modified data but the model used the same architecture and training process as our best-performing model. While the scores were lower in most cases (with the exception of walk prediction), they were still relatively close to the original scores. This shows that the model was able to learn the meaning of various patterns in the CGM signal rather than simply relying on the time of day. The second experiment ("No CGM") evaluated how much

Table 4.3: Ablation study results. Results of the modified models were significantly lower ($p < 0.05$) in nearly all cases when compared to our model, with the only exception being walk prediction with randomized time.

Model	Sleep AUROC	HR Fat Burn AUROC	HR Cardio AUROC	Walk AUROC
Our model	**0.947**	**0.722**	**0.768**	**0.817**
	(±0.001)	(±0.001)	(±0.005)	(±0.003)
Linear model	0.917	0.636	0.614	0.743
	(±0.001)	(±0.001)	(±0.004)	(±0.004)
Randomized time	0.891	0.690	0.691	**0.817**
	(±0.001)	(±0.002)	(±0.006)	(±0.003)
No CGM	0.911	0.703	0.701	0.713
	(±0.001)	(±0.002)	(±0.005)	(±0.004)
Mean ICD	0.944	0.713	0.756	0.811
	(±0.001)	(±0.002)	(±0.004)	(±0.003)
Non-pretrained ICD	0.945	0.706	0.709	0.813
	(±0.001)	(±0.002)	(±0.004)	(±0.003)

information is contained in the time of day when the CGM signal is not available. As in the previous experiment, the model architecture remained the same, so the model was aware of the medical claims and demographics and the only difference was that the CGM signal was replaced with zeros in both training and evaluation. Results for this experiment were slightly higher than for the previous experiment (with the exception of walk prediction) but again significantly lower than our proposed model. These results show that the time of day can be very useful for the predictions, and both time of day and CGM signal are needed to achieve the best results.

The last two experiments evaluated different ways of incorporating medical claims into the model. The first experiment ("Mean ICD") evaluated whether medical claims should be incorporated using an attention layer or if they could be combined by taking the average of the embedding vectors. The results show that using an attention layer improved the scores slightly, which means that the model was able to learn which ICD codes were more relevant for the predictions. However, using the mean embedding still provided improvements over a model which did not use ICD codes at all. The last experiment ("Non-pretrained ICD") evaluated the effect of using pretrained ICD embedding vectors. In this experiment, ICD embeddings were trained from scratch during the model training process. According to the results, training ICD embeddings during model training did not provide as good results as using pretrained embeddings. This is most likely due to the limited amount of training data, as most ICD codes appear relatively infrequently in our dataset. If more data were available, this approach could potentially reach better results than using pretrained embeddings because it would allow the model to distinguish between codes that have a similar description but which have different effects on the prediction targets. For example, the model could learn that a hand injury and a leg injury have different effects on walking.

4.4 Discussion

In this chapter, we have shown for the first time using a large cohort of people with T2D that CGM data is highly predictive of sleep and moderately predictive of physical activity or certain indicators of physical activity, as inferred by an activity tracker. To do so, we developed a neural network architecture for predicting daily activities using CGM, demographics, and claims data. We first modified the U-Net architecture so that it can be used with 1-dimensional CGM signals, and then incorporated claims data using an attention layer which allowed the model to learn which claims were important for the task. We showed that physical activity predictions improved when demographics and medical claims were incorpo-

rated into the model which we hypothesize to be because they can inform the model as to the presence of conditions which may affect heart rate, or they can show that the individual had physical restrictions (e.g., a leg injury) which lower the overall probability of being physically active. Sleep prediction did not benefit from the additional information.

A limitation of our study is that the data used for ground truth is collected by an activity tracker as opposed to the gold standard measurements, such as polysomnography. Since activity tracker data may sometimes be an inaccurate predictor of sleep or physical activities, our model may capture what is perceived by the activity tracker as sleep or physical activity rather than the true activities. Thus, comparing CGM to gold standard measurements remains an important direction for future work. Our work should also be taken into account in the context of people with T2D, and future work is needed to understand whether the results are generalizable to the entire population. For example, it is not unusual for individuals with diabetes to be on calcium channel blockers, beta blockers, or other medications that may blunt the heart rate response to activity. Another limitation of our study is that many individuals were mostly sedentary. For example, approximately half of them never walked for 10 consecutive minutes at a moderate pace and 10% of them never reached the fat burn zone. This limited the amount of data that could be used as positive examples of physical activity in both model training and evaluation.

In our future work, we will investigate using alternative approaches to determine physical activity, such as combining heart rate and walking into a more informative label, as neither of these labels is a perfect measure of physical activity on its own. For example, walking is not the only type of physical activity for many people and it might have a similar CGM pattern as other physical activities. On the other hand, estimated heart rate zones can be imprecise for some of the individuals. In addition, we will explore incorporating other types of data, such as medications, into the model. We expect these changes to improve our physical activity predictions. Another potential future direction is to personalize the models. For example, models could be fine-tuned for individuals by keeping track of their

daily activities for a short period of time before relying on the predictions from the model. This could be done by either wearing a fitness watch temporarily or by keeping track of the daily activities manually. Heart rate zones could also be improved by asking individuals to perform a short, moderate-intensity walk to determine what their heart rate is during a moderate physical activity.

4.5 Conclusion

In this chapter, we have presented a novel neural network architecture that can determine daily activities, including sleeping and physical activities, using CGM data while also considering other relevant information with varying time frequencies, such as medical history and demographics. Our results indicate that CGM data is highly predictive of sleep and has the potential to provide accurate predictions of physical activities if more precise labels were available. These results open new opportunities for the utilization of CGM devices in remote health monitoring for people with T2D.

CHAPTER 5

Identifying Substance Use and High-Risk Sexual Behavior Using Mobile Phone Data

5.1 Introduction

Men who have sex with men (MSM) are at higher risk of substance use and sexually transmitted infections (STIs) than the general population. For example, MSM are twice as likely to use illicit drugs [119], which may be used to cope with negative life events and thoughts, or to enhance pleasure during sex [22]. In addition, over half of new human immunodeficiency virus (HIV) infections occur among this population, which can be attributed to high-risk sexual behaviors and intravenous drug use (IDU) [67, 83]. Research suggests that these health disparities in substance use and HIV are generated by unjust social conditions [138, 154] and increased exposure to minority stressors [25, 121]. MSM are also at higher odds of mental distress and depression [75], which in turn may increase substance use as a coping mechanism [22].

Systematic reviews of intervention studies tailored for MSM have shown that interventions can be effective on methamphetamine- and sexual health-related outcomes, such as having sex without a condom or under the influence of drugs [100], and participants find them useful for gaining new knowledge and skills [120]. In addition, participants find them useful for self-reflection [120], which may lead to behavior change. However, a global survey among substance-using MSM found that only 11% of respondents had access to substance use treatment programs and 5% participated in such a program [65]. In the US, only 6.5%

of people who needed substance use treatment received it in 2020 [162]. Majority of people who were determined to need treatment did not recognize the need themselves, but among the ones who wanted treatment, the main reasons for not receiving it were affordability due to the lack of health care coverage, not finding an appropriate program, and fear of others having a negative opinion of them. Mobile- and eHealth-based interventions could improve the accessibility of interventions, as they can be accessed regardless of geographic location. In addition, the added privacy of eHealth interventions may reduce the concerns about treatment affecting other people's perception about the participant.

Mobile and eHealth interventions may also open new opportunities for personalization through increased availability of data about participants. Prior studies have shown success in providing personalized HIV interventions to MSM and people using substances [60, 87, 163]. However, this personalization typically depends on participants reporting behaviors manually which may become tedious. For example, one such study asked participants to respond to either daily or biweekly surveys, which many participants reported to be too repetitive or frequent regardless of the frequency [168]. However, the group receiving daily surveys found them to be more useful than the group receiving biweekly surveys, as it better reflected the frequent changes in behavior. This indicates that there may be a benefit in continuous monitoring of behaviors, but the monitoring should not depend on receiving frequent input from the participant. Therefore, being able to automate some or all the behavior monitoring could reduce the burden for participants.

In this chapter, we investigate how machine learning techniques can help identify risky behaviors among MSM from passively sensed mobile phone data. Prior studies have predicted HIV risk using Twitter [191], electronic health record (EHR) [102,117], or smartphone survey data [185]. Similarly, substance use risk has been predicted using survey data about personal characteristics [95], cognitive test results [1], Instagram profile data [79], and social media posts [135]. To the best of our knowledge, this is the first study to predict risky behaviors among MSM using passively collected mobile phone data, which allows for frequent data

collection with minimal effort required from the participant.

We first develop a mobile sensing application which tracks participant's daily actions, such as their location, messaging, and app use. We then train machine learning models to detect risky behaviors from this data and evaluate their performance on predicting different behaviors. Lastly, we analyze how different risky or protective behaviors manifest in mobile phone data.

The main contributions of this chapter are:

1. Demonstrating how passively collected mobile phone data can be used for risk prediction and identifying limitations of this approach.

2. Evaluating which types of data should be collected to identify risky behavior by training machine learning models using different subsets of the data, as well as analyzing differences between participants' data.

3. Determining how accurately different behaviors can be identified from mobile phone data.

5.2 Methods

5.2.1 Participant Recruitment

Participants were recruited using a variety of methods, including online outreach, in-person outreach, and through referral. Online outreach consisted of targeted, paid advertising on social media sites and websites such as Facebook, Instagram, Grindr, and Craigslist. Additionally, study staff created study-specific social media accounts and posted about the study on Facebook, Instagram, and through specific online communities that engage with MSM who use substances. In-person outreach included distributing recruitment materials such as business cards, flyers, and palm cards at events and locations attended by MSM,

such as sexual health clinics, community centers, substance use treatment centers, non-profit organizations, retail businesses, and Gay Pride festivals.

To be eligible for the study, participants had to meet the following criteria:

1. Be age 18 to 29
2. Identify as a sexual or gender minority
3. Have had anal or oral sex in the past 3 months
4. Have used substances (such as alcohol, marijuana, poppers, methamphetamines, heroin, cocaine, ecstasy, etc.) in the past 3 months
5. Have had sex while using substances in the past 3 months
6. Have a negative or unknown HIV status
7. Have used a dating app to meet sexual and substance use partners in the past 3 months
8. Own a smartphone
9. Reside in the United States
10. Be willing to participate in a 12-month study
11. Be able to provide informed consent

Consent was obtained during the onboarding and enrollment process, which took place on Zoom for all participants. The consent document provided details of what types of data were collected by the data collection app and the survey instruments. The consent document also provided information on how the participant's data would be protected and stored. If the participant consented to participate, their agreement was recorded by the interviewer, and they were emailed a copy of the consent to keep for their records.

Participants received up to US $350 in e-gift cards or web-based payment (i.e., PayPal, Venmo, Cash App, or Amazon gift cards) for their participation. The payments corresponded to completion of study activities, and not remote/passive app data collection.

5.2.2 Data Collection App

We developed a mobile app called eWellness for Android [9] phones to collect data on participant's mobile phone use activities. The app was based on the Aware Framework [64], which has been used in numerous earlier eHealth studies, for example to predict depression and anxiety [125, 134], progression of Parkinson's disease [179], or alcohol use events [13]. We asked participants to install our app on their personal phones and to give it all the necessary permissions to collect data in the background. After installing the eWellness app, participants were asked to leave it running in the background to collect their keyboard and location data throughout the entire time they were participating in the study. In addition to collecting keyboard data when participants typed text, the app collected information on which app they were typing the text in or, if they were using a browser, which website they were on.

The app also contained a sexual behavior and substance use survey (shown in Figure 5.1a) which the participants were asked to fill when they joined the study and once every three months after that. The survey was adapted from the U.S. Centers for Disease Control and Prevention (CDC)'s HIV/Pre-exposure prophylaxis (PrEP) clinical practice guideline [66]. The survey contains questions on one's substance use and sexual behaviors, which yields a total score indicating one's perceived risk of HIV infection and PrEP eligibility. The questions and answer options were:

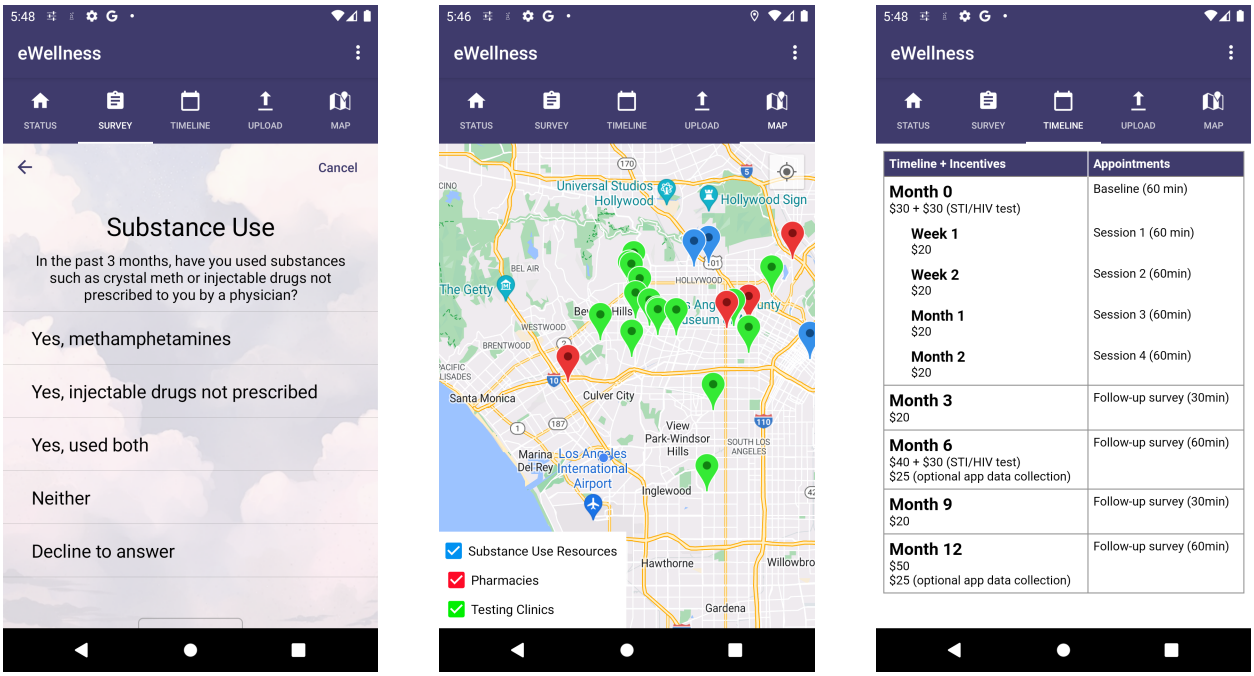
1. How old are you today? (Enter number)

2. In the past 3 months, have you used substances such as crystal meth or injectable drugs not prescribed to you by a physician? (Yes, methamphetamines; Yes, injectable

drugs not prescribed; Yes, used both; Neither; Decline to answer)

3. In the past 3 months, were you in a substance use in-patient or out-patient treatment program? (Yes; No; Decline to answer)
4. In the past 3 months, did you inject cocaine? (Yes; No; Decline to answer)
5. In the past 3 months, did you inject methamphetamines? (Yes; No; Decline to answer)
6. In the past 3 months, did you use needles? (Yes; No; Decline to answer)
7. In the past 3 months, did you share injection equipment? (Yes; No; Decline to answer)
8. In the past 3 months, did you inject in a group setting? (Yes; No; Decline to answer)
9. Do you currently take PrEP? PrEP stands for pre-exposure prophylaxis and it is a medication that helps prevent HIV transmission. (Yes; No; Decline to answer)
10. In the last 3 months, how many men have you had sex with? (Over 10; 6-10; 1-5; 0; Decline to answer)
11. In the last 3 months, how many times did you have receptive anal sex (you were the bottom) with a man when he did not use a condom? (1 or more times; 0 times; Decline to answer)
12. In the last 3 months, how many of your male sex partners were HIV-positive? (More than 1 HIV+ male partners; 1 HIV+ male partner; 0; Don't know; Decline to answer)
13. In the last 3 months, how many times did you have insertive anal sex (you were the top) with a man who was HIV-positive when you did not use a condom? (5 or more times; 0-4 times; Decline to answer)

To keep the participants engaged, the app also provided other useful information, such as a map of nearby resources, such as pharmacies, testing clinics, and substance use resources (Figure 5.1b), and the study timeline and incentives (Figure 5.1c).



(a) Sexual behavior and substance use survey

(b) Map of nearby resources, such as pharmacies and testing clinics

(c) Study timeline and incentives

Figure 5.1: Screenshots of the data collection app.

The app sent the collected data to our secure server every 30 minutes whenever internet connection was available. Highly sensitive information, such as passwords, were filtered out and the research team had no access to them. The server did not contain other identifying information, and only a randomly assigned participant identifier was included in the data. Mapping between personal information and participant identifiers was stored on separate computers and all the analysis was performed using only de-identified data.

5.2.3 Data Preprocessing

Keyboard Data

The eWellness app collected information about the currently active text field's contents on every keystroke. As a result, our database contained multiple rows of data for every full line of text that the participant wrote. For example, typing "Hello" might have been stored in the database as rows containing text values: "H", "He", "Hel", "Hell", and "Hello". In addition, the participant could have changed earlier parts of the text or used autocorrection, which means that this same text could have appeared as database rows: "H", "He", "Hel", "Helo", "Hello". As a result, the earlier row was not always a substring of the next one.

To remove duplicate rows, we repeated the following steps for each individual participant's text data until there were no more rows to remove:

1. Compare each row to the next row and if the first row is a substring of the second one, remove the first row.

2. Calculate the Levenshtein similarity between each row and the following row, and if the similarity is larger than 0.6, remove the first row.

Levenshtein similarity between two strings a and b is defined as:

$$sim(a,b) = 1 - \frac{dist(a,b)}{max(len(a),\ len(b))}$$

where dist(a,b) is the Levenshtein distance [107] which counts the minimum number of single-character modifications (insert, delete, substitute) that are necessary to make the strings identical:

$$dist(a,b) = \begin{cases} len(a), \text{ if } len(b) = 0, \text{ or vice versa} \\ dist(tail(a),\ tail(b)), \text{ if } a[0] = b[0] \\ 1 + min(dist(tail(a),\ b),\ dist(a,\ tail(b)),\ dist(tail(a),\ tail(b))) \end{cases}$$

where *a[0]* refers to the first character of string *a*, and *tail(a)* refers to a substring of *a* which contains everything except the first character. We calculated the similarity score using the TextDistance Python library [111].

App Data

For every row of text data, we had a package name of the app where the text was typed in as well as the Uniform Resource Locator (URL) of the website if the participant was using a web browser. In many cases, online services can be accessed both through an app and through a website, so we combined these data sources by extracting the domain name from the URL and mapping the commonly appearing URLs to the corresponding package names using a manually curated list of domain name - app pairs. If a domain name was not in this list, the domain name itself was used as the package name. Apps and websites were treated in the same manner in analysis and model training.

Location Data

The eWellness app saved GPS coordinates periodically whenever the phone moved to a different location. The app avoided unnecessary data collection to reduce battery consumption by only collecting location data when the phone was moving. This meant that if the participant remained in the same location, we did not receive location data until the participant started moving again. As we were only interested in locations where the participant spent time in rather than locations that the participants moved by, we removed data points where the participant was moving and only retained the last location once the movement had ended.

5.2.4 Feature Extraction

After cleaning the dataset using the previously shown steps, we manually extracted various features from each of the data sources to be used with the machine learning algorithms.

These feature extraction techniques are described in the following subsections.

Text Data

In our dataset, participants were active for different numbers of days, and for individual participants, different days had sometimes vastly different amounts of text data. This made the direct application of traditional text processing techniques challenging. In addition, the words and phrases used by MSM population sometimes differed from the ones used by the general public, so for optimal results, the techniques had to be tailored for this population. We only considered text data collected from social media, dating, or messaging apps/websites, as text data from other sources was found to contain more noise than useful information (for example, product names in shopping apps or location names in navigation apps).

The first set of features extracted from the text data was the frequencies of individual words used. Participants' text was first lemmatized, which means that inflected word forms were transformed to their base forms (e.g., "walking" -> "walk", "better" -> "good") to avoid having the same word appear in multiple forms in our set of features. Lemmatization was performed using WordNet Lemmatizer from Natural Language Toolkit (NLTK) [36]. Then, we removed words defined in NLTK's stop word list, commonly appearing placeholder texts (e.g., "Enter message", "Say something"), as well as words used by fewer than five people. For each remaining word, we calculated the frequency of word use:

$$freq(w) = \frac{\#\ days\ with\ word\ w}{\#\ days\ with\ text\ data}$$

The second set of text features only considered words and phrases associated with drug use or sexual behaviors. We used a phrase list which had been found effective for identifying MSM HIV risk behavior as well as substance use by an earlier study [135], and we used our previously defined frequency formula to determine frequencies for both individual phrases as well as for higher-level phrase categories (i.e., different types of substance use or sexual

behavior).

A third set of features was computed by Linguistic Inquiry and Word Count (LIWC) software [114], which uses built-in dictionaries to capture social and psychological states. It computes features describing how much an individual talks about a variety of topics, such as money, physical intimacy, or leisure activities, and it also computes higher-level descriptive features to measure factors, such as analytical thinking, authenticity, and emotional tone. It has been used in numerous studies to, for example, analyze fake news [193], social media posts [108], online reviews [190], and college admission essays [6]. We used it to generate features for each individual day, and we calculated the average across all days for each individual participant.

Our last set of text features was generated using the Bidirectional Encoder Representations from Transformers (BERT) language model [57]. We used a model that was pretrained for sentiment analysis using Twitter data [14], as we expected Twitter data to use similar language as other social media and messaging platforms. We removed the last fully connected layer of the model so that it could be used to generate text embeddings, and we applied it to each individual day of data. These text embeddings were then averaged across all days of data for individual participants.

App Data

We captured app usage by looking at apps where the user wrote text. We generated one set of features by calculating how frequently each app was used:

$$freq(app) = \frac{\#\ days\ using\ app}{\#\ days\ using\ any\ apps}$$

We also considered a subset of these frequency features which only contained social media, dating, and messaging apps, as we expected other types of apps to be less relevant to our prediction task. Other apps (e.g., maps, music, or shopping) were expected to be noisy and

therefore to have a negative effect on the model's predictive performance.

Location Data

Before extracting features from location data, we clustered GPS coordinates for each individual participant by using the Mean-Shift algorithm [44]. This algorithm moves all points repeatedly towards the mean value of their neighborhood (determined by window radius r) until all points have converged. Points that converge to the same coordinates are defined to belong in the same cluster, thus allowing the algorithm to find the appropriate number of clusters. As the generated clusters depend on the window size, we determined the appropriate size by visually inspecting the clustering results. We also assumed that the most visited location was the participant's home.

We then computed the features describing individual's mobility, such as how far from home they travelled, how many locations they visited per day, and how many of these locations were unique. These features were selected such that they were potentially related to participant's risk level either directly or indirectly. For example, number of unique locations may be associated with having many partners, while having very few unique locations could be related to methamphetamine use due to the limited number of locations where the participant could safely use it. The full list of location-based features is shown below:

1. Maximum/mean/median distance from home
2. Percentage of days spent over 10 or over 50 miles away from home
3. Percentage of days (or 2 or 3 consecutive days) spent entirely away from home
4. Percentage of days spent away from two or three of the most common locations
5. Percentage of days with more than one or two location(s)
6. Average number of locations visited

7. Number of unique locations visited

5.2.5 Model Training

We used Scikit-learn [141] to train logistic regression [54] and gradient boosting [68] classification models to predict participants' answers for each survey question. These models were chosen to represent a simple linear model as well as a more advanced non-linear model. To determine which types of data could be useful for the prediction task, we trained separate models using individual data categories, such as location data, app use, and risky word use. We then evaluated combinations of these features, focusing on feature combinations that we believed to give a comprehensive view to the participant's activities without including redundant data (e.g., not including social media apps and all apps in the same model). Models were evaluated using leave-one-out cross-validation due to the relatively small number of participants.

5.3 Results

5.3.1 Data Statistics

We collected data from 65 participants between November 25, 2020 - January 9, 2023. Dataset statistics are shown in Table 5.1. Among all the apps, we manually identified 66 social media, dating, and messaging apps which were later used to analyze participants' messaging data. It should be noted that apps include unique websites as well (grouped by domain name).

We used data for all participants who had at least 30 days of data available. If a participant's answer to a survey question was "Decline to answer" or "I don't know", this answer was not included in the model training or evaluation. This did not, however, exclude their other survey answers from being used. Statistics for survey responses are shown in Table 5.2

Table 5.1: Dataset statistics.

	Total	Mean	Median	SD	Min	Max
Lines of text	2,176,879	32,983.0	20,944	27,865.8	2,529	110,803
Locations	905,127	15,341.1	5,566	22,765.9	509	93,024
Unique apps	1,828	85.3	72	59.0	20.0	338.0
Age		25.4	26	3.4	18	41

(full questions and answer options are shown in Section 5.2.2).

5.3.2 Model Performance

We trained classification models to predict answers to each question. As some questions had partially overlapping answer options, we split them to multiple distinct questions and trained separate models for each of them. For example, the answer options for substance use included methamphetamine use, injectable drug use, and both, so we trained separate models for predicting methamphetamine use and injectable drug use. Individuals who responded that they used both were given a positive label for both prediction tasks. In addition, some questions had a very low number of positive responses (for example, only two participants were in a substance use treatment program) or were determined to not always indicate high-risk behaviors (for example, having a HIV+ partner), so the focus of our discussion will be on the five questions which we determined to be the most informative.

Results for predicting survey responses using both individual feature types as well as combinations of them are shown in Table 5.3. Feature combinations were selected both based on their individual results and based on whether they were presumed to provide non-overlapping information.

As the results show, methamphetamine use could be predicted well using just the text data. Multiple text-based approaches worked well, including LIWC, BERT, and word fre-

Table 5.2: Survey response statistics.

	Positive	Negative	Total
Methamphetamine use	22 (34%)	43 (66%)	65
Injectable drug use	8 (12%)	57 (88%)	65
Injects cocaine	3 (5%)	62 (95%)	65
Injects in group	4 (6%)	61 (94%)	65
Shares injection equipment	3 (5%)	62 (95%)	65
Injects methamphetamine	6 (9%)	59 (91%)	65
Condomless receptive sex	45 (69%)	20 (31%)	65
Condomless insertive sex w/ HIV+ partner 5+ times	3 (5%)	54 (95%)	57
HIV+ partners	6 (11%)	51 (89%)	57
Over 5 partners	34 (52%)	31 (48%)	65
Over 10 partners	21 (32%)	44 (68%)	65
Takes PrEP	31 (48%)	34 (52%)	65
In substance use treatment program	2 (3%)	63 (97%)	65

quency model. Combining multiple feature types improved the results only very slightly. Similarly, PrEP use could also be predicted well using either text data or location data, but combining multiple feature types improved the results. Predicting having many partners worked also reasonably well when combining all feature types. Predictive models were only moderately successful in determining whether the participant had condomless receptive sex or used injectable drugs, the latter of which had very few positive responses.

Combining multiple feature types rarely improved the performance by a noticeable amount. This could be because in many cases the feature groups might provide redundant information, so using only one highly informative feature group was enough. In addition, increasing the number of features could lead to overfitting, as the number of features can become much

Table 5.3: F1 scores for predicting answers to survey questions. F1 score was calculated for the less frequent response, which in most cases was the positive answer (answer frequencies are shown in Table 5.2). First value shows the score using logistic regression and the second value shows the score using a gradient boosting classifier.

	Social apps	All apps	Location	Risky words	All words	LIWC	BERT	Risky words, LIWC	Social apps, Risky words, LIWC	Location, Social apps, Risky words, LIWC	All
Substance use											
Methamphetamine use	.64 / .56	.63 / .59	.46 / .70	.67 / .72	.75 / .54	.84 / .42	.84 / .68	.84 / .61	.84 / .75	**.86** / .53	**.86** / .44
Injectable drug use	.00 / .15	.10 / **.56**	.31 / .12	.14/ .00	.00 / .00	.46 / .38	.29 / .47	.27 / .25	.29 / .25	.00 / .00	.00 / .00
Sexual behavior											
6+ partners	.37 / .47	.47 / .47	.54 / .51	.65 / .41	.51 / .53	.45 / .46	.48 / .60	.51 / .23	.56 / .39	.62 / .52	.62 / **.72**
Condomless receptive sex	.27 / .18	.42 / .29	.37 / .31	.29 / .25	.22 / .36	.41 / .35	.41 / .35	.32 / .21	**.45** / .19	.40 / .07	.32 / .44
Protective behaviors											
Takes PrEP	.58 / .63	.57 / .67	.63 / .74	.56 / .56	.68 / .60	.49 / .58	.57 / .58	.62 / .58	.59 / .56	.71 / .48	.68 / **.76**

larger than the number of participants.

5.3.3 Feature Analysis

Next, we analyzed how the participant data differed depending on the survey responses. We show the differences for the most predictive tasks, which were methamphetamine use, taking PrEP, and having 6+ partners.

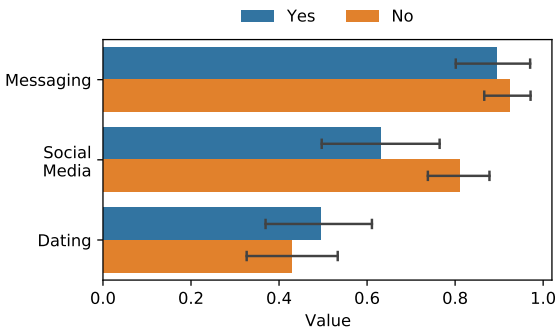
App Use

Figure 5.2 shows how frequently participants used apps from different categories. We considered any apps that are used for communicating with other people and divided them into three categories: messaging apps (e.g., Messages, WhatsApp, Telegram), social media apps (e.g., Facebook, Instagram, Reddit), and dating apps (e.g., Tinder, Grindr, Adam4Adam).

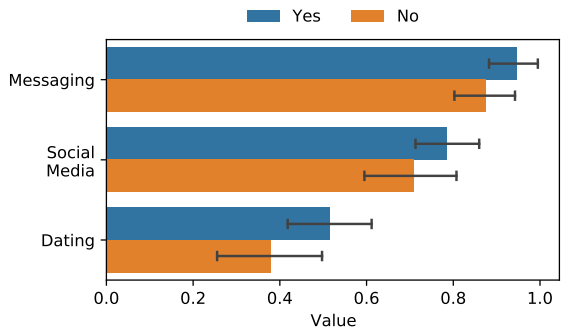
Methamphetamine users were less likely to use social media apps than non-users (p=0.01). Participants who had 6+ partners or who used PrEP were consistently more active in all social app categories, but these differences were not statistically significant. The largest difference in both cases was in the use of dating apps (p=0.09 and 0.06 respectively).

Risky Words

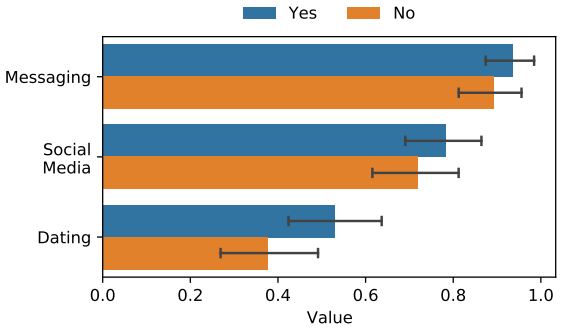
Figure 5.3 shows differences in risky word use. We divided the list of risky words to sex-related and drug-related words. Methamphetamine users had a higher occurrence of sex-related words, but due to the relatively small number of methamphetamine users, this difference was not statistically significant (p=0.19). There was only a very minor (statistically insignificant) difference in the frequency of drug word use. People with 6+ partners and PrEP users were both more likely to use sex-related words (p=0.04 and 0.08 respectively).



(a) Methamphetamine use

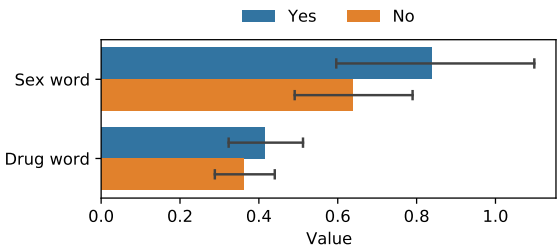


(b) Having 6+ partners

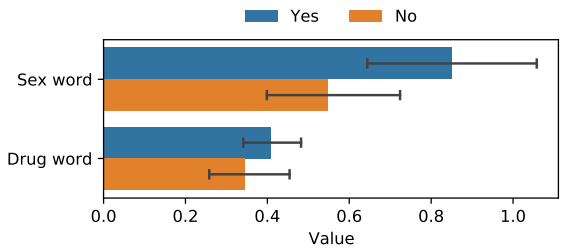


(c) PrEP use

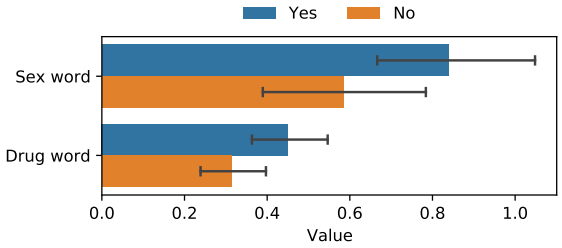
Figure 5.2: Differences in app use among different groups. X-axis represents the percentage of days when the participant communicated using an app from a certain category.



(a) Methamphetamine use



(b) Having 6+ partners



(c) PrEP use

Figure 5.3: Differences in risky word use among different groups. X-axis represents the percentage of days when the participant used words or phrases from a certain category.

Location

Figure 5.4 shows how location data differed for different groups. Due to the large number of location-based features, we chose a smaller subset of features which contained less overlapping information. Methamphetamine users were less likely to spend time over 50 miles from home (p=0.02) and their trips were not as far away (p=0.05). In addition, they visited slightly fewer distinct locations, although the difference was not statistically significant (p=0.15). People having 6+ partners were more active overall, visiting more locations (p=0.003), traveling further (p=0.03), and spending multiple days in a row away from home more frequently (p=0.05). PrEP users were also more likely to visit more locations (p=0.001), and they were more likely to travel far away from home (p=0.001).

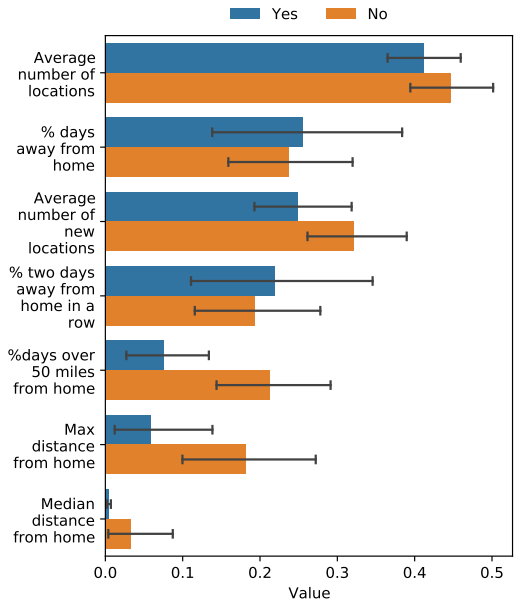
LIWC

Lastly, in Figure 5.5, we compare LIWC features among different groups. Again, due to the large number of distinct features, we show results only for some of the super-categories which we expected to show differences. Methamphetamine users were more likely to use social (p<0.001) and affect words (p=0.007) and less likely to use drive related words (p=0.006). People having 6 or more partners were slightly more likely to use drive related words (p=0.03), and PrEP users were more likely to use cognitive process words (p=0.009).

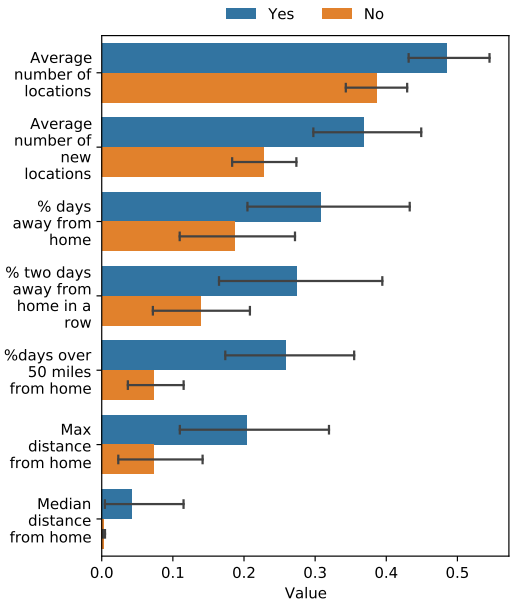
5.4 Discussion

5.4.1 Principal Results

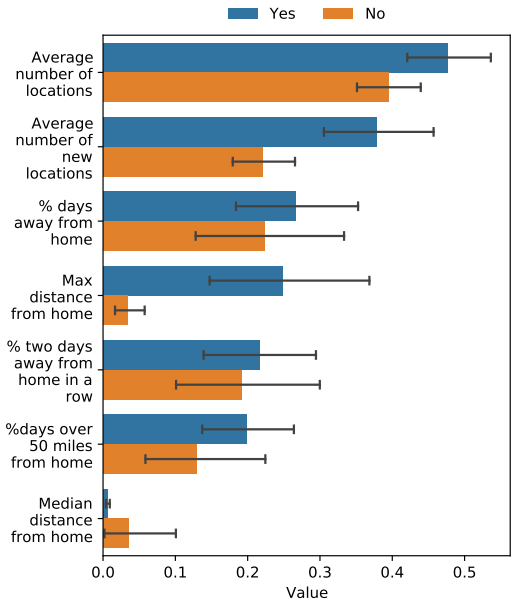
In this chapter, we have shown that mobile sensing data can be used to identify multiple risky behaviors as well as preventative measures taken by the individuals. More specifically, our participants' text and location data were highly informative of methamphetamine use, taking PrEP, and having many sexual partners. In addition, the results were promising for



(a) Methamphetamine use

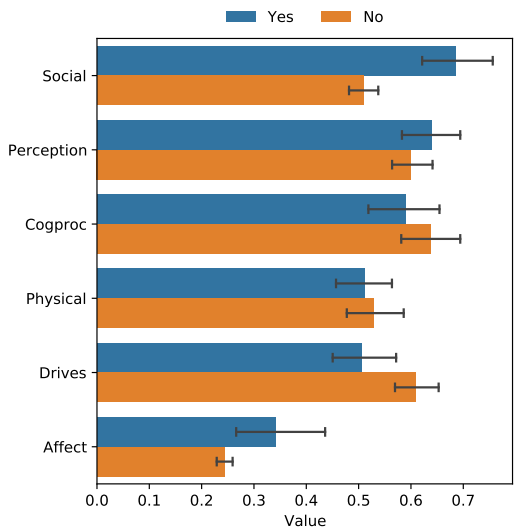


(b) Having 6+ partners

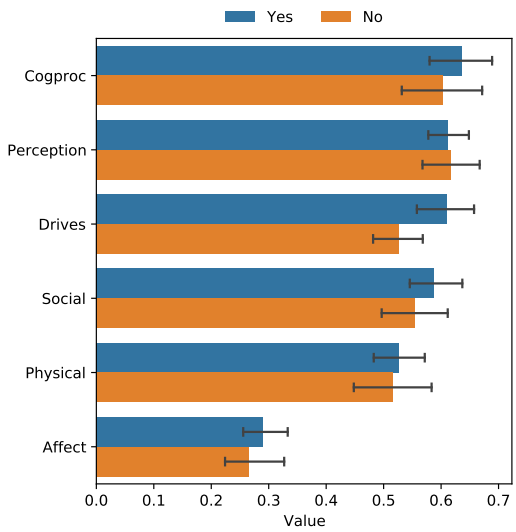


(c) PrEP use

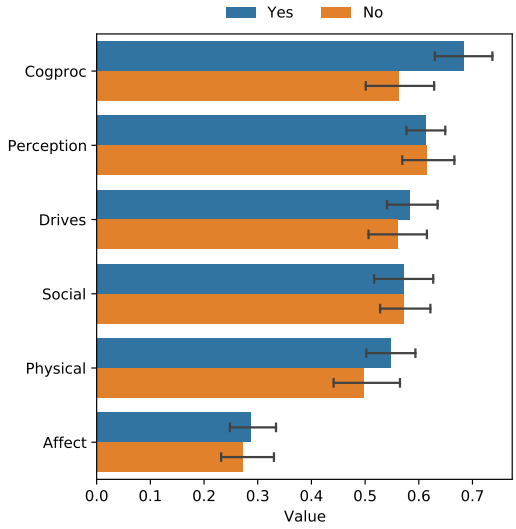
Figure 5.4: Differences in location data among different groups. Values have been scaled such that the largest individual value for each feature becomes 1 to be able to show all values in the same figure.



(a) Methamphetamine use



(b) Having 6+ partners



(c) PrEP use

Figure 5.5: Differences in LIWC features among different groups. Original values have been scaled to fit in the same figure.

some of the less common risky behaviors, such as injectable drug use, and more data may help improve the results further. As some of these behaviors are quite infrequent among the MSM population, a more targeted participant recruitment may be needed to gather sufficient amount of data.

In addition to determining which behaviors can be predicted, our second goal was to determine what data is useful for these predictions. We have shown that text-based features were the most informative for most behaviors, which was an expected result because participants might, for example, look for partners on dating apps or discuss substance use in private messages with other people. This matches with the findings of [135], which showed that certain types of substance use may be predicted from social media messaging data. In most cases, collecting all text data was not necessary and better results were achieved by utilizing pre-determined word lists, such as LIWC or risky word list defined in [135]. Therefore, an intervention app may improve its privacy by collecting a more limited set of messaging data.

In addition, we have shown that more recent language modeling techniques, such as BERT, can often provide similar results as the traditional techniques based on pre-determined word lists and word frequencies. However, the more abstract nature of these representations may complicate the interpretation of the results, as individual values do not have a human-interpretable meaning. On the other hand, BERT representations can also help improve privacy, as they do not reveal which exact words the participants used. Due to the small number of participants, training a language model using our dataset was not feasible, so we relied on a model that had been trained on Twitter data. While we expect the language use to be mostly similar across datasets, training a language model using data from the target population could improve the results if enough data were available, as people might use different words and phrases in Twitter compared to dating apps or private messages.

We were also able to detect behavioral differences between groups of people with different survey responses. For example, methamphetamine users were more likely to use sex-related

words in their messages. Earlier research has shown that methamphetamine users have more sexual partners [70] and may be engaged in more risky sexual behavior, such as unprotected sex, although it is not clear whether methamphetamine use is causing the behavior [26]. This increased sexual activity is likely also causing an increase in sex-related discussions. Methamphetamine users were also less likely to travel far from home. This may be related to their lower household income level [96], which could make traveling far unaffordable. They also used more affective and social words and fewer drive-related words, which may be partly related to the higher prevalence of co-occurring mental health problems [96].

We also found that people with many sexual partners were more active users of all types of social apps (messaging, social media, dating). Earlier studies have shown that users of geosocial networking apps, such as Grindr, have more sexual partners in general [72, 84], and MSM with more partners have larger social networks [156], which may explain the more frequent use of social apps. Being more social may similarly explain more time spent away from home and in many different locations. People with many sexual partners also used more sex- and drug-related words. Substance use has previously been found to be associated with a larger number of sexual partners [31].

PrEP users were found to be more active users of dating apps, which shows that MSM who were looking for sexual partners were also more likely to take precautions against the risks. This finding is in line with an earlier study which showed that Grindr users were more likely to take PrEP than MSM who did not use the app [84]. We also found PrEP users to be more likely to use both sex- and drug-related words. Substance use may be related to PrEP use because substance users are more likely to have more sexual partners [84] and have unprotected sex [21], which increases their HIV risk. In addition, injectable drug use also increases HIV risk and therefore users are more likely to have a prescription for PrEP [160]. However, relatively few participants in our study reported using injectable drugs. PrEP users were also found to travel further from home and visit more distinct locations. This may be related to PrEP users having a higher income level on average and therefore having

the means to travel more [28,164]. Their LIWC score for cognitive processes was also higher which indicates higher complexity in writing [170]. This may be due to PrEP users being older on average, as the youngest adults are less likely to use PrEP than older adults relative to their risk level [28,157].

5.4.2 Comparison with Prior Work

The closest work similar to ours is [135] which identified HIV as well as amphetamine, methamphetamine and tetrahydrocannabinol (THC) use from social media messaging data. Our work differs from this by using a wider range of data sources collected through participants' mobile devices. For example, our study used text typed in any mobile app and website which allowed us to identify risky behaviors in traditional messaging apps and less common dating apps in addition to the most popular social media apps. In addition, we utilized location data, which allows us to analyze participants' daily movement patterns and their relation to risky behaviors. We also attempted to identify a wider range of risky behaviors, especially related to sexual health. The only shared prediction target between these papers was methamphetamine use, which the earlier paper was able to predict with F1 score of 0.85. This is very close to our result (0.86), which provides support for these findings.

Another similar study is [79] which predicted alcohol, tobacco, prescription drug, and illegal drug use from Instagram data. They were able to detect alcohol use with statistical significance, but they had less success in predicting other types of substance use. Our better prediction results may be attributed to having access to more personal messaging data, as many people may avoid discussing substance use on public platforms. This shows that choosing the appropriate data collection methods is very important for accurate results.

Other studies have implemented personalized MSM interventions using survey data [15], identified how effective MSM-targeted mobile app interventions are [189], or evaluated the feasibility and acceptability of mobile sensing among MSM population [61,62,174]. However, these studies have not evaluated whether mobile sensing data can be used to inform and

personalize interventions, which was the goal of our study.

5.4.3 Limitations

One limitation of our study was that we only included participants who had an Android smartphone. We chose to only include Android users because iPhones have more restrictions on what data can be collected, and therefore collecting text data would have been unfeasible. Many potential participants had to be excluded from the study because of their mobile device, which may skew the demographics to some extent. In addition, as the data was collected using personal devices, there were some interruptions in data collection. For example, some participants turned off or deleted the app during the study while others upgraded to a new phone without re-installing the app. In addition, some Android phones were found to have a rather aggressive battery saving functionality which occasionally turned off the data collection. To avoid data collection issues, we kept track of when each participant's device had last sent us data and contacted participants after a few missing days to make sure the data collection would be resumed.

Another limitation was that the text data only included what the participants typed on their phones. This approach may miss the context of some messages, as the responses were not collected. It could be informative to know what content participants consumed online or what messages they received from others. In addition, participants might have messaged with people using multiple devices (e.g., a computer or a tablet in addition to their phone), so our data collection approach might not have been able to track all social media usage and messaging for some participants.

Lastly, some of the outcomes that we set out to predict were very infrequent which made the task impossible. For example, only two of our participants were in a substance use treatment program, which was not enough for training and evaluating a machine learning model. Therefore, we had to focus on questions which had a reasonable number of both positive and negative responses.

5.4.4 Future Work

Our future work will explore providing personalized interventions using predictive models to determine which types of interventions may be appropriate. We will, for example, investigate sending participants resources that may help them in their current life situation, such as providing information about PrEP to individuals who may be at elevated HIV risk based on their substance use or sexual behavior but who are not yet taking it.

5.5 Conclusion

In this chapter, we have presented a mobile sensing application which can identify certain types of substance use, high-risk sexual behavior, and protective actions from passsively collected smartphone data. We have shown this data to be highly predictive of methamphetamine use, having many sexual partners, or taking PrEP. While further work is still needed to evaluate how effective interventions based on automatic behavior tracking are, these results show that mobile sensing applications could be used to personalize interventions for high-risk individuals. This can reduce the burden of participating in intervention programs, as the daily behaviors can be tracked with minimal effort from the participants.

CHAPTER 6

Conclusion

In this dissertation, we have presented algorithmic and technological solutions for providing personalized healthcare in both hospital and remote settings. In a hospital setting, our proposed algorithms aim to assist medical professionals in making decisions more efficiently. Specifically, we proposed algorithms that can adaptively choose the order of medical tests to perform accurate diagnosis at a low cost, and we also proposed an algorithm that can assist in segmenting medical images which reduces the radiologist's workload.

However, healthcare is not limited to hospital care. Therefore, we also presented solutions for remote health monitoring, which enables personalized care throughout individuals' daily lives and may allow for earlier interventions. Specifically, we demonstrated how continuous glucose monitors can be used to determine individuals' daily activities to inform diabetes management and interventions, and we explored how mobile phone data can be used to identify individuals at high risk of substance use or high-risk sexual activities so that they can receive targeted interventions.

Given the wide range of healthcare-related issues, there is a need for further research to propose new solutions to other healthcare problems. A special focus should be placed on preventative measures that can identify future health problems before they become difficult to treat. This can be achieved through the use of personal sensors such as mobile phones, smart watches, or specialized physiological sensors which can be used to track individuals, or by analyzing medical history, including previously diagnosed health problems, laboratory test results, and medications, to identify individuals at high risk of future health problems.

REFERENCES

[1] Woo-Young Ahn, Divya Ramesh, Frederick Gerard Moeller, and Jasmin Vassileva. Utility of machine-learning approaches to identify behavioral markers for substance use disorders: impulsivity dimensions as predictors of current cocaine dependence. *Frontiers in psychiatry*, 7:34, 2016.

[2] Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. Machine-learning-based disease diagnosis: A comprehensive review. In *Healthcare*, volume 10, page 541. MDPI, 2022.

[3] Mahmoud Al-Ayyoub, Duaa Alawad, Khaldun Al-Darabsah, and Inad Aljarrah. Automatic detection and classification of brain hemorrhages. *WSEAS transactions on computers*, 12(10):395–405, 2013.

[4] Duaa Mohammad Alawad, Avdesh Mishra, and Md Tamjidul Hoque. Aibh: Accurate identification of brain hemorrhage using genetic algorithm based feature selection and stacking. *Machine Learning and Knowledge Extraction*, 2(2):56–77, 2020.

[5] Roohallah Alizadehsani, Moloud Abdar, Mohamad Roshanzamir, Abbas Khosravi, Parham M Kebria, Fahime Khozeimeh, Saeid Nahavandi, Nizal Sarrafzadegan, and U Rajendra Acharya. Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in biology and medicine*, 111:103346, 2019.

[6] AJ Alvero, Sonia Giebel, Ben Gebre-Medhin, Anthony Lising Antonio, Mitchell L Stevens, and Benjamin W Domingue. Essay content and style are strongly related to household income and sat scores: Evidence from 60,000 undergraduate applications. *Science advances*, 7(42):eabi9031, 2021.

[7] American Diabetes Association Professional Practice Committee. 1. Improving care and promoting health in populations: Standards of medical care in diabetes - 2022. *Diabetes Care*, 45(Supplement 1):S8–S16, 12 2021. doi:10.2337/dc22-S001.

[8] American Diabetes Association Professional Practice Committee. 5. Facilitating behavior change and well-being to improve health outcomes: Standards of medical care in diabetes - 2022. *Diabetes Care*, 45(Supplement 1):S60–S82, 12 2021. doi:10.2337/dc22-S005.

[9] Android. The platform pushing what's possible, 2022. Accessed October 5, 2022. URL: http://www.android.com.

[10] Ali Arab, Betty Chinda, George Medvedev, William Siu, Hui Guo, Tao Gu, Sylvain Moreno, Ghassan Hamarneh, Martin Ester, and Xiaowei Song. A fast and fully-automated deep-learning approach for accurate hemorrhage segmentation and volume quantification in non-contrast whole-head ct. *Scientific Reports*, 10(1):1–12, 2020.

[11] Mohammad Reza Askari, Mudassir Rashid, Xiaoyu Sun, Mert Sevil, Andrew Shahidehpour, Keigo Kawaji, and Ali Cinar. Meal and physical activity detection from free-living data for discovering disturbance patterns of glucose levels in people with diabetes. *BioMedInformatics*, 2(2):297–317, 2022. doi:10.3390/biomedinformatics2020019.

[12] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7), 2015. doi:10.1371/journal.pone.0130140.

[13] Sangwon Bae, Tammy Chung, Denzil Ferreira, Anind K Dey, and Brian Suffoletto. Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. *Addictive behaviors*, 83:42–47, 2018.

[14] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.

[15] José A Bauermeister, Emily S Pingel, Laura Jadwin-Cakmak, Gary W Harper, Keith Horvath, Gretchen Weiss, and Patricia Dittus. Acceptability and preliminary efficacy of a tailored online hiv/sti testing intervention for young men who have sex with men: the get connected! program. *AIDS and Behavior*, 19:1860–1874, 2015.

[16] Vladimir B. Berikov, Olga A. Kutnenko, Julia F. Semenova, and Vadim V. Klimontov. Machine learning models for nocturnal hypoglycemia prediction in hospitalized patients with type 1 diabetes. *Journal of Personalized Medicine*, 12(8), 2022. doi:10.3390/jpm12081262.

[17] Lauriane Bertrand, Nathan Cleyet-Marrel, and Zilu Liang. The role of continuous glucose monitoring in automatic detection of eating activities. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 313–314, 2021. doi:10.1109/LifeTech52111.2021.9391849.

[18] H. S. Bhadauria and M. L. Dewal. Intracranial hemorrhage detection using spatial fuzzy c-mean and region-based active contour on brain CT imaging. *Signal, Image and Video Processing*, 8(2):357–364, 2014.

[19] H. S. Bhadauria, Annapurna Singh, and M. L. Dewal. An integrated method for hemorrhage segmentation from brain CT Imaging. *Computers and Electrical Engineering*, 2013. doi:10.1016/j.compeleceng.2013.04.010.

[20] Peter Boersma, Lindsey I Black, and Brian W Ward. Peer reviewed: prevalence of multiple chronic conditions among us adults, 2018. *Preventing chronic disease*, 17, 2020.

[21] Melissa R Boone, Stephanie H Cook, and Patrick Wilson. Substance use and sexual risk behavior in hiv-positive men who have sex with men: an episode-level analysis. *AIDS and Behavior*, 17:1883–1887, 2013.

[22] Adam Bourne and Peter Weatherburn. Substance use among men who have sex with men: patterns, motivations, impacts and intervention development need. *Sexually transmitted infections*, 93(5):342–346, 2017.

[23] Frédéric Bousefsaf, Djamaleddine Djeldjli, Yassine Ouzar, Choubeila Maaoui, and Alain Pruski. iPPG 2 cPPG: Reconstructing contact from imaging photoplethysmographic signals using U-Net architectures. *Computers in Biology and Medicine*, 138:104860, 2021.

[24] Joseph P Broderick, Thomas G Brott, John E Duldner, Thomas Tomsick, and Gertrude Huster. Volume of intracerebral hemorrhage. a powerful and easy-to-use predictor of 30-day mortality. *Stroke*, 24(7):987–993, 1993.

[25] Virginia R Brooks. *Minority stress and lesbian women*. Free Press, 1981.

[26] Joanne Bryant, Max Hopwood, Gary W Dowsett, Peter Aggleton, Martin Holt, Toby Lea, Kerryn Drysdale, and Carla Treloar. The rush to risk when interrogating the relationship between methamphetamine use and sexual practice among gay and bisexual men. *International Journal of Drug Policy*, 55:242–248, 2018.

[27] B. Barla Cambazoglu, Hugo Zaragoza, Olivier Chapelle, Jiang Chen, Ciya Liao, Zhaohui Zheng, and Jon Degenhardt. Early exit optimizations for additive machine learned ranking systems. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 411–420. ACM, 2010.

[28] Mitchell Caponi, Carolyne Burgess, Alexandra Leatherwood, and Luis Freddy Molano. Demographic characteristics associated with the use of HIV pre-exposure prophylaxis (prep) in an urban, community health center. *Preventive Medicine Reports*, 15:100889, 2019.

[29] Isabella Castiglioni, Leonardo Rundo, Marina Codari, Giovanni Di Leo, Christian Salvatore, Matteo Interlenghi, Francesca Gallivanone, Andrea Cozzi, Natascha Claudia D'Amico, and Francesco Sardanelli. AI applications to medical images: From machine learning to deep learning. *Physica Medica*, 83:9–24, 2021. doi:https://doi.org/10.1016/j.ejmp.2021.02.006.

[30] Guillaume Cathelain, Bertrand Rivet, Sophie Achard, Jean Bergounioux, and François Jouen. U-net neural network for heartbeat detection in ballistocardiography. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 465–468. IEEE, 2020.

[31] Patricia A Cavazos-Rehg, Melissa J Krauss, Edward L Spitznagel, Mario Schootman, Linda B Cottler, and Laura Jean Bierut. Number of sexual partners and associations with initiation and intensity of substance use. *AIDS and Behavior*, 15:869–874, 2011.

[32] Center for Devices and Radiological Health. FDA's role in regulating medical devices. URL: https://www.fda.gov/medical-devices/home-use-devices/fdas-role-regulating-medical-devices.

[33] Centers for Disease Control and Prevention. Diabetes home. https://www.cdc.gov/diabetes/basics/getting-tested.html, Aug 2017. Accessed July 23, 2018. URL: https://www.cdc.gov/diabetes/basics/getting-tested.html.

[34] Centers for Disease Control and Prevention. Questionnaires, datasets, and related documentation. https://wwwn.cdc.gov/nchs/nhanes/Default.aspx, 2018. Accessed July 23, 2018. URL: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx.

[35] Centers for Disease Control and Prevention. Type 2 diabetes, Dec 2021. Accessed September 9, 2022. URL: https://www.cdc.gov/diabetes/basics/type2.html.

[36] Centers for Disease Control and Prevention. Prevent diabetes complications, Mar 2022. Accessed January 11, 2023. URL: https://www.cdc.gov/diabetes/managing/problems.html.

[37] Centers for Medicaid and Medicare Services. National health expenditure data. Accessed January 11, 2023. URL: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.

[38] Michael Chan, Venu G. Ganti, and Omer T. Inan. Respiratory rate estimation using U-Net-based cascaded framework from electrocardiogram and seismocardiogram signals. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2481–2492, 2022. doi:10.1109/JBHI.2022.3144990.

[39] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers and Electrical Engineering*, 2014. doi:10.1016/j.compeleceng.2013.11.024.

[40] Wenbing Chang, Yinglai Liu, Yiyong Xiao, Xinglong Yuan, Xingxing Xu, Siyue Zhang, and Shenghan Zhou. A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics*, 9(4):178, 2019.

[41] Olivier Chapelle. Yahoo! Learning to Rank Challenge Overview. *JMLR: Workshop and Conference Proceedings*, 14:1–24, 2011.

[42] Minmin Chen, Zhixiang Xu, Kilian Q. Weinberger, Olivier Chapelle, and Dor Kedem. Classifier Cascade for Minimizing Feature Evaluation Cost. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[43] Suming Chen, Arthur Choi, and Adnan Darwiche. Computer adaptive testing using the same-decision probability. In *CEUR Workshop Proceedings*, 2015.

[44] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

[45] Guillaume Chevance, Natalie M Golaszewski, Elizabeth Tipton, Eric B Hekler, Matthew Buman, Gregory J Welk, Kevin Patrick, and Job G Godino. Accuracy and precision of energy expenditure, heart rate, and steps measured by combined-sensing fitbits against reference measures: Systematic review and meta-analysis. *JMIR Mhealth Uhealth*, 10(4):e35626, Apr 2022. doi:10.2196/35626.

[46] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Development and validation of deep learning algorithms for detection of critical findings in head CT scans, 2018. arXiv:1803.05854.

[47] Jungrae Cho, Inchul Choi, Jaeil Kim, Sungmoon Jeong, Young-Sup Lee, Jaechan Park, Jungjoon Kim, and Minho Lee. Affinity graph based end-to-end deep convolutional networks for ct hemorrhage segmentation. In *International Conference on Neural Information Processing*, pages 546–555. Springer, 2019.

[48] Arthur Choi, Yexiang Xue, and Adnan Darwiche. Same-decision probability: A confidence measure for threshold-based decisions. In *International Journal of Approximate Reasoning*, 2012. doi:10.1016/j.ijar.2012.04.005.

[49] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. Phase-aware speech enhancement with deep complex u-net. In *International Conference on Learning Representations*, 2018.

[50] Y. Choi, A. Darwiche, and G. Van Den Broeck. Optimal feature selection for decision robustness in Bayesian networks. In *IJCAI International Joint Conference on Artificial Intelligence*, 2017.

[51] Chui S Chu, Nikki P Lee, John Adeoye, Peter Thomson, and Siu-Wai Choi. Machine learning and treatment outcome prediction for oral cancer. *Journal of Oral Pathology & Medicine*, 49(10):977–985, 2020.

[52] Keh Shih Chuang, Hong Long Tzeng, Sharon Chen, Jay Wu, and Tzong Jer Chen. Fuzzy c-means clustering with spatial information for image segmentation. *Computerized Medical Imaging and Graphics*, 2006. doi:10.1016/j.compmedimag.2005.10.001.

[53] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation.

In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[54] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[55] Darpit Dave, Daniel J. DeSalvo, Balakrishna Haridas, Siripoom McKay, Akhil Shenoy, Chester J. Koh, Mark Lawley, and Madhav Erraguntla. Feature-based machine learning model for real-time hypoglycemia prediction. *Journal of Diabetes Science and Technology*, 15(4):842–855, Jul 2021. doi:10.1177/1932296820922622.

[56] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[57] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[58] DICOM. Current edition. Accessed February 28, 2021. URL: https://www.dicomstandard.org/current/.

[59] Joseph L. Dieleman, Ellen Squires, Anthony L. Bui, Madeline Campbell, Abigail Chapin, Hannah Hamavid, Cody Horst, Zhiyin Li, Taylor Matyasz, Alex Reynolds, Nafis Sadat, Matthew T. Schneider, and Christopher J. L. Murray. Factors Associated With Increases in US Health Care Spending, 1996-2013. *JAMA*, 318(17):1668–1678, 11 2017. doi:10.1001/jama.2017.15927.

[60] Rebecca Dillingham, Karen Ingersoll, Tabor E Flickinger, Ava Lena Waldman, Marika Grabowski, Colleen Laurence, Erin Wispelwey, George Reynolds, Mark Conaway, and Wendy F Cohn. Positivelinks: a mobile health intervention for retention in hiv care and clinical outcomes with 12-month follow-up. *AIDS patient care and STDs*, 32(6):241–250, 2018.

[61] Dustin T Duncan, Basile Chaix, Seann D Regan, Su Hyun Park, Cordarian Draper, William C Goedel, June A Gipson, Vincent Guilamo-Ramos, Perry N Halkitis, Russell Brewer, et al. Collecting mobility data with gps methods to understand the hiv environmental riskscape among young black men who have sex with men: A multi-city feasibility study in the deep south. *AIDS and Behavior*, 22:3057–3070, 2018.

[62] Dustin T Duncan, Farzana Kapadia, Seann D Regan, William C Goedel, Michael D Levy, Staci C Barton, Samuel R Friedman, and Perry N Halkitis. Feasibility and acceptability of global positioning system (gps) methods to study the spatial contexts of substance use and sexual risk behaviors among young men who have sex with men in new york city: A p18 cohort sub-study. *PloS one*, 11(2):e0147520, 2016.

[63] Kirstin Early, Stephen E. Fienberg, and Jennifer Mankoff. Test time feature ordering with FOCUS. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, pages 992–1003, 2016. doi:10.1145/2971648.2971748.

[64] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[65] Juan M Flores, Glenn-Milo Santos, Keletso Makofane, Sonya Arreola, and George Ayala. Availability and use of substance abuse treatment programs among substance-using men who have sex with men worldwide. *Substance use & misuse*, 52(5):666–673, 2017.

[66] Centers for Disease Control and Prevention. Preexposure prophylaxis for the prevention of HIV infection in the United States—2017 update: a clinical practice guideline, 2017. Accessed February 14, 2023. URL: https://www.cdc.gov/hiv/pdf/risk/prep/cdc-hiv-prep-guidelines-2017.pdf.

[67] Centers for Disease Control and Prevention. HIV surveillance report, 2023. Accessed February 14, 2023. URL: http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html.

[68] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[69] Anjali Gautam and Balasubramanian Raman. Automatic segmentation of intracerebral hemorrhage from brain CT images. In M. Tanveer and Ram Bilas Pachori, editors, *Machine Intelligence and Signal Analysis*, pages 753–764, Singapore, 2019. Springer Singapore.

[70] David R Gibson, Martin H Leamon, and Neil Flynn. Epidemiology and public health consequences of methamphetamine use in California's Central Valley. *Journal of psychoactive drugs*, 34(3):313–319, 2002.

[71] Céline R. Gillebert, Glyn W. Humphreys, and Dante Mantini. Automated delineation of stroke lesions using brain CT images. *NeuroImage: Clinical*, 4:540 – 548, 2014. doi:https://doi.org/10.1016/j.nicl.2014.03.009.

[72] William C Goedel and Dustin T Duncan. Geosocial-networking app usage patterns of gay, bisexual, and other men who have sex with men: Survey among users of Grindr, a mobile dating app. *JMIR public health and surveillance*, 1(1):e4353, 2015.

[73] Sara Bersche Golas, Takuma Shibahara, Stephen Agboola, Hiroko Otaki, Jumpei Sato, Tatsuya Nakae, Toru Hisamitsu, Go Kojima, Jennifer Felsted, Sujay Kakarmath, et al. A machine learning model to predict the risk of 30-day readmissions in patients with

heart failure: a retrospective analysis of electronic medical records data. *BMC medical informatics and decision making*, 18(1):1–17, 2018.

[74] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13).

[75] Gilbert Gonzales and Carrie Henning-Smith. Health disparities by sexual orientation: results and implications from the behavioral risk factor surveillance system. *Journal of community health*, 42:1163–1172, 2017.

[76] Monika Grewal, Muktabh Mayank Srivastava, Pulkit Kumar, and Srikrishna Varadarajan. Radnet: Radiologist level accuracy using deep learning for hemorrhage detection in ct scans. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 281–284. IEEE, 2018.

[77] Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. Accuracy of wristband fitbit models in assessing sleep: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 21(11):e16273, Nov 2019. doi:10.2196/16273.

[78] K. Haselsberger, R. Pucher, and L. M. Auer. Prognosis after acute subdural or epidural haemorrhage. *Acta Neurochirurgica*, 90(3-4):111–116, 1988. doi:10.1007/bf01560563.

[79] Saeed Hassanpour, Naofumi Tomita, Timothy DeLise, Benjamin Crosier, and Lisa A Marsch. Identifying substance use risk based on deep neural networks and instagram social media data. *Neuropsychopharmacology*, 44(3):487–494, 2019.

[80] Amir Hayeri. Predicting Future Glucose Fluctuations Using Machine Learning and Wearable Sensor Data. *Diabetes*, 67(Supplement 1), 07 2018. 738-P. doi:10.2337/db18-738-P.

[81] Jeremy J. Heit, Michael Iv, and Max Wintermark. Imaging of intracranial hemorrhage. *Journal of Stroke*, 19(1):11–27, 2017. doi:10.5853/jos.2016.00563.

[82] Brian L Hill, Nadav Rakocz, Ákos Rudas, Jeffrey N Chiang, Sidong Wang, Ira Hofer, Maxime Cannesson, and Eran Halperin. Imputation of the continuous arterial line blood pressure waveform from non-invasive measurements using deep learning. *Scientific reports*, 11(1):1–12, 2021.

[83] HIV.gov. Who is at risk for HIV?, 2023. Accessed February 14, 2023. URL: https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/who-is-at-risk-for-hiv.

[84] Martin Hoenigl, Susan J Little, David Grelotti, Britt Skaathun, Gabriel A Wagner, Nadir Weibel, Jamila K Stockman, and Davey M Smith. Grindr users take more risks, but are more open to human immunodeficiency virus (hiv) pre-exposure prophylaxis: could this dating app provide a platform for HIV prevention outreach? *Clinical Infectious Diseases*, 71(7):e135–e140, 2020.

[85] Murtadha Hssayeni. Computed tomography images for intracranial hemorrhage detection and segmentation, Mar 2020. URL: https://doi.org/10.13026/w8q8-ky94.

[86] Murtadha D. Hssayeni, Muayad S. Croock, Aymen D. Salman, Hassan Falah Al-khafaji, Zakaria A. Yahya, and Behnaz Ghoraani. Intracranial hemorrhage segmentation using a deep convolutional model. *Data*, 5(1), 2020. doi:10.3390/data5010014.

[87] Karen S Ingersoll, Rebecca A Dillingham, Jennifer E Hettema, Mark Conaway, Jason Freeman, George Reynolds, and Sharzad Hosseinbor. Pilot rct of bidirectional text messaging for ART adherence among nonurban substance users with HIV. *Health Psychology*, 34(S):1305, 2015.

[88] Natasha Ironside, Ching-Jen Chen, Simukayi Mutasa, Justin L Sim, Saurabh Marfatia, David Roh, Dale Ding, Stephan A Mayer, Angela Lignelli, and Edward Sander Connolly. Fully automated segmentation algorithm for hematoma volumetric analysis in spontaneous intracerebral hemorrhage. *Stroke*, 50(12):3416–3423, 2019.

[89] Mobarakol Islam, Parita Sanghani, Angela An Qi See, Michael Lucas James, Nicolas Kon Kam King, and Hongliang Ren. Ichnet: Intracerebral hemorrhage (ich) segmentation using deep learning. In *International MICCAI Brainlesion Workshop*, pages 456–463. Springer, 2018.

[90] Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. Classification with costly features using deep reinforcement learning. *arXiv preprint arXiv:1711.07364*, 2017.

[91] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-net convolutional networks. *ISMIR Conference*, 2017.

[92] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. URL: http://portal.acm.org/citation.cfm?doid=582415.582418, doi:10.1145/582415.582418.

[93] Fan Jia, Wing Hong Wong, and Tieyong Zeng. Ddunet: Dense dense U-net with applications in image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 354–364, 2021.

[94] Yazan Jian, Michel Pasquier, Assim Sagahyroon, and Fadi Aloul. A machine learning approach to predicting diabetes complications. *Healthcare*, 9(12):1712, December 2021. doi:10.3390/healthcare9121712.

[95] Yankang Jing, Ziheng Hu, Peihao Fan, Ying Xue, Lirong Wang, Ralph E Tarter, Levent Kirisci, Junmei Wang, Michael Vanyukov, and Xiang-Qun Xie. Analysis of substance use and its outcomes by machine learning I. Childhood evaluation of liability to substance use disorder. *Drug and alcohol dependence*, 206:107605, 2020.

[96] Christopher M Jones, Wilson M Compton, and Desiree Mustaquim. Patterns and characteristics of methamphetamine use among adults—united states, 2015–2018. *Morbidity and Mortality Weekly Report*, 69(12):317, 2020.

[97] Mohammad Kachuee, Sajad Darabi, Babak Moatamed, and Majid Sarrafzadeh. Dynamic feature acquisition using denoising autoencoders. *IEEE transactions on neural networks and learning systems*, 2018.

[98] Mohammad Kachuee, Orpaz Goldstein, Kimmo Kärkkäinen, Sajad Darabi, and Majid Sarrafzadeh. Opportunistic learning: Budgeted cost-sensitive learning from data streams. In *International Conference on Learning Representations*, 2019.

[99] Kaggle. RSNA intracranial hemorrhage detection. Accessed November 20, 2019. URL: https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/overview.

[100] Rod Knight, Mohammad Karamouzian, Anna Carson, Joshua Edward, Patrizia Carrieri, Jean Shoveller, Nadia Fairbairn, Evan Wood, and Danya Fast. Interventions to address substance use and sexual risk among gay, bisexual and other men who have sex with men who use methamphetamine: a systematic review. *Drug and Alcohol Dependence*, 194:410–429, 2019.

[101] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[102] Douglas S Krakower, Susan Gruber, Katherine Hsu, John T Menchaca, Judith C Maro, Benjamin A Kruskal, Ira B Wilson, Kenneth H Mayer, and Michael Klompas. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. *The Lancet HIV*, 6(10):e696–e704, 2019.

[103] Rita V Krishnamurthi, Valery L Feigin, Mohammad H Forouzanfar, George A Mensah, Myles Connor, Derrick A Bennett, Andrew E Moran, Ralph L Sacco, Laurie M Anderson, Thomas Truelsen, Martin O'Donnell, Narayanaswamy Venketasubramanian,

Suzanne Barker-Collo, Carlene M M Lawes, Wenzhi Wang, Yukito Shinohara, Emma Witt, Majid Ezzati, Mohsen Naghavi, and Christopher Murray. Global and regional burden of first-ever ischaemic and haemorrhagic stroke during 1990–2010: findings from the global burden of disease study 2010. *The Lancet Global Health*, 1(5):e259 – e281, 2013. doi:https://doi.org/10.1016/S2214-109X(13)70089-5.

[104] Matt J. Kusner, Wenlin Chen, Quan Zhou, Zhixiang Xu, Kilian Q. Weinberger, and Yixin Chen. Feature-cost sensitive learning with submodular trees of classifiers. *Proceedings of the National Conference on Artificial Intelligence*, 2014.

[105] Yeong Chan Lee, Sang-Hyuk Jung, Aman Kumar, Injeong Shim, Minku Song, Min Seo Kim, Kyunga Kim, Woojae Myung, Woong-Yang Park, and Hong-Hee Won. Icd2vec: Mathematical representation of diseases. *Journal of Biomedical Informatics*, page 104361, 2023.

[106] Rachel Leproult, Gaétane Deliens, Médhi Gilson, and Philippe Peigneux. Beneficial Impact of Sleep Extension on Fasting Insulin Sensitivity in Adults with Habitual Sleep Restriction. *Sleep*, 38(5):707–715, 05 2015. doi:10.5665/sleep.4660.

[107] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[108] Lifang Li, Qingpeng Zhang, Xiao Wang, Jun Zhang, Tao Wang, Tian-Lu Gao, Wei Duan, Kelvin Kam-fai Tsoi, and Fei-Yue Wang. Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. *IEEE Transactions on computational social systems*, 7(2):556–562, 2020.

[109] Lu Li, Meng Wei, Bo Liu, Kunakorn Atchaneeyasakul, Fugen Zhou, Zehao Pan, Shimran Kumar, Jason Zhang, Yuehua Pu, David Sigmund Liebeskind, et al. Deep learning for hemorrhagic lesion detection and segmentation on brain CT images. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[110] Chun Chih Liao, Furen Xiao, Jau Min Wong, and I. Jen Chiang. Computer-aided diagnosis of intracranial hematoma with brain deformation on computed tomography. *Computerized Medical Imaging and Graphics*, 2010. doi:10.1016/j.compmedimag.2010.03.003.

[111] Life4. Life4/Textdistance: Compute distance between sequences., 2022. Accessed October 5, 2022. URL: https://github.com/life4/textdistance.

[112] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 773–782, 2018.

[113] Wei Liu, Qiong Yan, and Yuzhi Zhao. Densely self-guided wavelet network for image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 432–433, 2020.

[114] LIWC. Welcome to LIWC-22, 2022. Accessed October 5, 2022. URL: https://www.liwc.app/.

[115] Sven Loncaric, Atam P. Dhawan, Joseph Broderick, and Thomas Brott. 3-d image analysis of intra-cerebral brain hemorrhage from digitized ct films. *Computer Methods and Programs in Biomedicine*, 46(3):207 – 216, 1995. doi:https://doi.org/10.1016/0169-2607(95)01620-9.

[116] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[117] Julia L Marcus, Leo B Hurley, Douglas S Krakower, Stacey Alexeeff, Michael J Silverberg, and Jonathan E Volk. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *The lancet HIV*, 6(10):e688–e695, 2019.

[118] Yonit Marcus, Roy Eldor, Mariana Yaron, Sigal Shaklai, Maya Ish-Shalom, Gabi Shefer, Naftali Stern, Nehor Golan, Amit Z. Dvir, Ofir Pele, and Mira Gonen. Improving blood glucose level predictability using machine learning. *Diabetes/Metabolism Research and Reviews*, 36(8):e3348, 2020. doi:https://doi.org/10.1002/dmrr.3348.

[119] GLRN Medley, Rachel N Lipari, Jonaki Bose, Devon S Cribb, Larry A Kroutil, Gretchen McHenry, et al. Sexual orientation and estimates of adult substance use and mental health: Results from the 2015 national survey on drug use and health. *NSDUH data review*, 10:1–54, 2016.

[120] Rebecca Meiksin, GJ Melendez-Torres, Jane Falconer, T Charles Witzel, Peter Weatherburn, and Chris Bonell. ehealth interventions to address sexual health, substance use, and mental health among men who have sex with men: systematic review and synthesis of process evaluations. *Journal of medical Internet research*, 23(4):e22477, 2021.

[121] Ilan H Meyer. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. *Psychological bulletin*, 129(5):674, 2003.

[122] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[123] Xu Min, Bin Yu, and Fei Wang. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on copd. *Scientific reports*, 9(1):1–10, 2019.

[124] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 2017. doi:10.1016/j.patcog.2016.11.008.

[125] Isaac Moshe, Yannik Terhorst, Kennedy Opoku Asare, Lasse Bosse Sander, Denzil Ferreira, Harald Baumeister, David C. Mohr, and Laura Pulkki-Råback. Predicting symptoms of depression and anxiety using smartphone and wearable data. *Frontiers in Psychiatry*, 12, 2021. doi:10.3389/fpsyt.2021.625247.

[126] Arif Muhammad and Wang Guojun. Segmentation of calcification and brain hemorrhage with midline detection. In *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*, pages 1082–1090. IEEE, 2017.

[127] John Muschelli, Elizabeth M Sweeney, Natalie L Ullman, Paul Vespa, Daniel F Hanley, and Ciprian M Crainiceanu. Pitchperfect: Primary intracranial hemorrhage probability estimation using random forests on ct. *NeuroImage: Clinical*, 14:379–390, 2017.

[128] Feng Nan and Venkatesh Saligrama. Adaptive classification for prediction under a budget. In *Advances in Neural Information Processing Systems*, pages 4727–4737, 2017.

[129] Feng Nan, Joseph Wang, and Venkatesh Saligrama. Feature-Budgeted Random Forest. *Proceedings of The 32nd International Conference on Machine Learning*, 37:1983–1991, 2015.

[130] Feng Nan, Joseph Wang, and Venkatesh Saligrama. Pruning random forests for prediction on a budget. In *Advances in neural information processing systems*, pages 2334–2342, 2016.

[131] National Institute of Diabetes and Digestive and Kidney Diseases. Symptoms & causes of diabetes, 2022. Accessed January 11, 2023. URL: https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes.

[132] Giulia Noaro, Giacomo Cappon, Martina Vettoretti, Giovanni Sparacino, Simone Del Favero, and Andrea Facchinetti. Machine-learning based model to improve insulin bolus calculation in type 1 diabetes therapy. *IEEE Transactions on Biomedical Engineering*, 68(1):247–255, 2021. doi:10.1109/TBME.2020.3004031.

[133] National Institutes of Health. Nifti-1 data format - neuroimaging informatics technology initiative. https://nifti.nimh.nih.gov/nifti-1. Accessed February 28, 2021. URL: https://nifti.nimh.nih.gov/nifti-1.

[134] Kennedy Opoku Asare, Yannik Terhorst, Julio Vega, Ella Peltonen, Eemil Lagerspetz, and Denzil Ferreira. Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study. *JMIR mHealth and uHealth*, 9(7):e26540, 2021.

[135] Anaelia Ovalle, Orpaz Goldstein, Mohammad Kachuee, Elizabeth SC Wu, Chenglin Hong, Ian W Holloway, and Majid Sarrafzadeh. Leveraging social media activity and machine learning for hiv and substance abuse risk assessment: development and validation study. *Journal of Medical Internet Research*, 23(4):e22042, 2021.

[136] Victor Palacios, Diane Myung-Kyung Woodbridge, and Jean L. Fry. Machine learning-based meal detection using continuous glucose monitoring on healthy participants: An objective measure of participant compliance to protocol. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 7032–7035, 2021. doi:10.1109/EMBC46164.2021.9630408.

[137] Bumjun Park, Songhyun Yu, and Jechang Jeong. Densely connected hierarchical network for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019.

[138] Jeffrey T Parsons, H Jonathon Rendina, Raymond L Moody, Ana Ventuneac, and Christian Grov. Syndemic production and sexual compulsivity/hypersexuality in highly sexually active gay and bisexual men: Further evidence for a three group conceptualization. *Archives of sexual behavior*, 44:1903–1913, 2015.

[139] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative adversarial network. *Proc. Interspeech 2017*, pages 3642–3646, 2017.

[140] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[141] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[142] Murali Gundu Rao, Dalbir Singh, Niranjan Khandelwal, and Suresh Kumar Sharma. Dating of early subdural haematoma: a correlative clinico-radiological study. *Journal of clinical and diagnostic research: JCDR*, 10(4):HC01, 2016.

[143] Soumi Ray, Vinod Kumar, Chirag Ahuja, and Niranjan Khandelwal. Intensity population based unsupervised hemorrhage segmentation from brain CT images. *Expert Systems with Applications*, 97:325 – 335, 2018. doi:https://doi.org/10.1016/j.eswa.2017.12.032.

[144] Eduardo Pontes Reis, Felipe Nascimento, Mateus Aranha, Fernando Mainetti Secol, Birajara Machado, Marcelo Felix, Anouk Stein, and Edson Amaro. Brain hemorrhage extended (bhx): Bounding box extrapolation from thick to thin slice ct images. URL: https://physionet.org/content/bhx-brain-bounding-box/1.1/, doi:10.13026/9CFT-HG92.

[145] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015.

[146] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[147] Saima Safdar, Saad Zafar, Nadeem Zafar, and Naurin Farooq Khan. Machine learning based decision support systems (dss) for heart disease diagnosis: a review. *Artificial Intelligence Review*, 50(4):597–623, 2018.

[148] Frank A. J. L. Scheer, Michael F. Hilton, Christos S. Mantzoros, and Steven A. Shea. Adverse metabolic and cardiovascular consequences of circadian misalignment. *Proceedings of the National Academy of Sciences*, 106(11):4453–4458, 2009. doi:10.1073/pnas.0808180106.

[149] Moritz Scherer, Jonas Cordes, Alexander Younsi, Yasemin-Aylin Sahin, Michael Götz, Markus Möhlenbruch, Christian Stock, Julian Bösel, Andreas Unterberg, Klaus Maier-Hein, et al. Development and validation of an automatic segmentation algorithm for quantification of intracerebral hemorrhage. *Stroke*, 47(11):2776–2782, 2016.

[150] Wonju Seo, You-Bin Lee, Seunghyun Lee, Sang-Man Jin, and Sung-Min Park. A machine-learning approach to predict postprandial hypoglycemia. *BMC Medical Informatics and Decision Making*, 19(1):210, Dec 2019. doi:10.1186/s12911-019-0943-4.

[151] Khader Shameer, Kipp W Johnson, Alexandre Yahi, Riccardo Miotto, LI Li, Doran Ricks, Jebakumar Jebakaran, Patricia Kovatch, Partho P Sengupta, Sengupta Gelijns, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using mount sinai heart failure cohort. In *Pacific Symposium on Biocomputing 2017*, pages 276–287. World Scientific, 2017.

[152] Bhavna Sharma and K Venugopalan. Automatic segmentation of brain CT scan image to identify hemorrhages. *International Journal of Computer Applications*, 40(10):0975–8887, 2012.

[153] Matthew F Sharrock, W Andrew Mould, Hasan Ali, Meghan Hildreth, Issam A Awad, Daniel F Hanley, and John Muschelli. 3D Deep Neural Network Segmentation of Intracerebral Hemorrhage: Development and Validation for Clinical Trials. *Neuroinformatics*, 2020. doi:10.1007/s12021-020-09493-5.

[154] Merill Singer and Scott Clair. Syndemics and public health: Reconceptualizing disease in bio-social context. *Medical anthropology quarterly*, 17(4):423–441, 2003.

[155] J. Slaght, M. Sénéchal, T. J. Hrubeniuk, A. Mayo, and D. R. Bouchard. Walking cadence to exercise at moderate intensity for adults: A systematic review. *Journal of Sports Medicine*, 2017:1–12, 2017. doi:10.1155/2017/4641203.

[156] AMA Smith, Jeffrey Grierson, David Wain, Marian Pitts, and Pip Pattison. Associations between the sexual behaviour of men who have sex with men and the structure and composition of their social networks. *Sexually transmitted infections*, 80(6):455–458, 2004.

[157] Jonathan M Snowden, Yea-Hung Chen, Willi McFarland, and Henry F Raymond. Prevalence and characteristics of users of pre-exposure prophylaxis (prep) among men who have sex with men, san francisco, 2014 in a cross-sectional survey: implications for disparities. *Sexually transmitted infections*, 93(1):52–55, 2017.

[158] Karine Spiegel, Rachel Leproult, and Eve Van Cauter. Impact of sleep debt on metabolic and endocrine function. *The lancet*, 354(9188):1435–1439, 1999.

[159] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *ISMIR Conference*, 2018.

[160] Carl G Streed, Jake R Morgan, Mam Jarra Gai, Marc R Larochelle, Michael K Paasche-Orlow, and Jessica L Taylor. Prevalence of HIV preexposure prophylaxis prescribing among persons with commercial insurance and likely injection drug use. *JAMA Network Open*, 5(7):e2221346–e2221346, 2022.

[161] Joel Stremmel, Brian L Hill, Jeffrey Hertzberg, Jaime Murillo, Llewelyn Allotey, and Eran Halperin. Extend and explain: Interpreting very long language models. In *Machine Learning for Health*, pages 218–258. PMLR, 2022.

[162] Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the united states: Results from the 2020 national survey on drug use and health. *HHS Publication No. PEP20-07-01-001, NSDUH Series H-55*, 2021.

[163] Patrick S Sullivan, Robert Driggers, Joanne D Stekler, Aaron Siegler, Tamar Goldenberg, Sarah J McDougal, Jason Caucutt, Jeb Jones, and Rob Stephenson. Usability and acceptability of a mobile comprehensive HIV prevention app for men who have sex with men: a pilot study. *JMIR mHealth and uHealth*, 5(3):e7199, 2017.

[164] Patrick Sean Sullivan, Cory Woodyatt, Chelsea Koski, Elizabeth Pembleton, Pema McGuinness, Jennifer Taussig, Alexandra Ricca, Nicole Luisi, Eve Mokotoff, Nanette Benbow, et al. A data visualization and dissemination resource to support HIV prevention and care at the local level: analysis and uses of the AIDSVu public data resource. *Journal of medical Internet research*, 22(10):e23173, 2020.

[165] Deepak Sumbria, Engin Berber, Manikannan Mathayan, and Barry T Rouse. Virus infections and host metabolism—can we manage the interactions? *Frontiers in Immunology*, 11:594963, 2021.

[166] Hong Sun, Pouya Saeedi, Suvi Karuranga, Moritz Pinkepank, Katherine Ogurtsova, Bruce B. Duncan, Caroline Stein, Abdul Basit, Juliana C.N. Chan, Jean Claude Mbanya, Meda E. Pavkov, Ambady Ramachandaran, Sarah H. Wild, Steven James, William H. Herman, Ping Zhang, Christian Bommer, Shihchen Kuo, Edward J. Boyko, and Dianna J. Magliano. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, 183:109119, 2022. doi:https://doi.org/10.1016/j.diabres.2021.109119.

[167] M. Sun, R. Hu, H. Yu, B. Zhao, and H. Ren. Intracranial hemorrhage detection by 3d voxel segmentation on brain CT images. In *2015 International Conference on Wireless Communications Signal Processing (WCSP)*, pages 1–5, 2015. doi:10.1109/WCSP.2015.7341238.

[168] Dallas Swendeman, Nithya Ramanathan, Laura Baetscher, Melissa Medich, Aaron Scheffler, W Scott Comulada, and Deborah Estrin. Smartphone self-monitoring to support self-management among people living with HIV: Perceived benefits and theory of change from a mixed-methods, randomized pilot study. *Journal of acquired immune deficiency syndromes (1999)*, 69(0 1):S80, 2015.

[169] Naoya Takahashi, Nabarun Goswami, and Yuki Mitsufuji. MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 106–110, 2018. doi:10.1109/IWAENC.2018.8521383.

[170] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[171] The Diabetes Prevention Program (DPP) Research Group. The Diabetes Prevention Program (DPP): Description of lifestyle intervention. *Diabetes Care*, 25(12):2165–2171, 12 2002. doi:10.2337/diacare.25.12.2165.

[172] Daniel N. Thyde, Ali Mohebbi, Henrik Bengtsson, Morten Lind Jensen, and Morten Mørup. Machine learning-based adherence detection of type 2 diabetes patients

on once-daily basal insulin injections. *Journal of Diabetes Science and Technology*, 15(1):98–108, 2021. PMID: 32297804. doi:10.1177/1932296820912411.

[173] Lucas R Trambaiolli, Ana C Lorena, Francisco J Fraga, Paulo AM Kanda, Renato Anghinah, and Ricardo Nitrini. Improving alzheimer's disease diagnosis with machine learning techniques. *Clinical EEG and neuroscience*, 42(3):160–165, 2011.

[174] Kathy Trang, Patrick S Sullivan, Devon E Hinton, Carol M Worthman, Minh Giang Le, and Tanja Jovanovic. Feasibility, acceptability, and design of a mobile health application for high-risk men who have sex with men in hanoi, vietnam. *The Lancet Global Health*, 8:S14, 2020.

[175] Kirill Trapeznikov, Venkatesh Saligrama, and David Castañón. Multi-stage classifier design. *Machine Learning*, 92(2-3):479–502, 9 2013. doi:10.1007/s10994-013-5349-4.

[176] Riku Turkki, Dmitrii Byckhov, Mikael Lundin, Jorma Isola, Stig Nordling, Panu E Kovanen, Clare Verrill, Karl von Smitten, Heikki Joensuu, Johan Lundin, et al. Breast cancer outcome prediction with tumour tissue images and machine learning. *Breast cancer research and treatment*, 177(1):41–52, 2019.

[177] William P. T. M. van Doorn, Yuri D. Foreman, Nicolaas C. Schaper, Hans H. C. M. Savelberg, Annemarie Koster, Carla J. H. van der Kallen, Anke Wesselius, Miranda T. Schram, Ronald M. A. Henry, Pieter C. Dagnelie, Bastiaan E. de Galan, Otto Bekers, Coen D. A. Stehouwer, Steven J. R. Meex, and Martijn C. G. J. Brouwers. Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The maastricht study. *PLOS ONE*, 16(6):1–17, 06 2021. doi:10.1371/journal.pone.0253125.

[178] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[179] Julio Vega. Monitoring parkinson's disease progression using behavioural inferences, mobile devices and web technologies. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 323–327, 2016.

[180] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[181] Joseph Wang. An LP for Sequential Learning Under Budgets. *AISTATS*, 2014. doi:10.1007/978-3-319-10593-2{_}39.

[182] Joseph Wang, Kirill Trapeznikov, and Venkatesh Saligrama. Efficient Learning by Directed Acyclic Graph For Resource Constrained Prediction. *Advances in Neural Information Processing Systems*, 2015.

[183] Weichen Wang, Shayan Mirjafari, Gabriella Harari, Dror Ben-Zeev, Rachel Brian, Tanzeem Choudhury, Marta Hauser, John Kane, Kizito Masaba, Subigya Nepal, et al. Social sensing: assessing social functioning of patients living with schizophrenia using mobile phone sensing. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15, 2020.

[184] World Health Organization. International Classification of Diseases (ICD). Accessed September 20, 2022. URL: https://www.who.int/standards/classifications/classification-of-diseases.

[185] Tyler B Wray, Xi Luo, Jun Ke, Ashley E Pérez, Daniel J Carr, and Peter M Monti. Using smartphone survey data and machine learning to identify situational and contextual risk factors for hiv risk behavior among men who have sex with men who are not on prep. *Prevention Science*, 20:904–913, 2019.

[186] Yuan Xie, Bin Jiang, Enhao Gong, Ying Li, Guangming Zhu, Patrik Michel, Max Wintermark, and Greg Zaharchuk. Use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. *American Journal of Roentgenology*, 212(1):44–51, 2019.

[187] Zhixiang Xu, Matt Kusner, Kilian Weinberger, and Minmin Chen. Cost-sensitive tree of classifiers. In *International Conference on Machine Learning*, pages 133–141, 2013.

[188] Zhixiang Eddie Xu, Kilian Q Weinberger, Olivier Chapelle, St Louis, and Olivier Chapelle Cc. The Greedy Miser: Learning under Test-time Budgets. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1175–1182, 2012.

[189] Jin Yan, Aidi Zhang, Liang Zhou, Zhulin Huang, Pan Zhang, and Guoli Yang. Development and effectiveness of a mobile phone application conducting health behavioral intervention among men who have sex with men, a randomized controlled trial: study protocol. *BMC Public Health*, 17(1):1–8, 2017.

[190] Dezhi Yin, Samuel D Bond, and Han Zhang. Anxious or angry? effects of discrete emotions on the perceived helpfulness of online reviews. *MIS quarterly*, 38(2):539–560, 2014.

[191] Sean D Young, Wenchao Yu, and Wei Wang. Toward automating hiv identification: machine learning for rapid identification of hiv-related social media data. *Journal of acquired immune deficiency syndromes (1999)*, 74(Suppl 2):S128, 2017.

[192] David M. Yousem and Rohini Nadgir. *Neuroradiology: the requisites*. Elsevier, 2017.

[193] Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, 2020.