# Poster: Identifying Causal Patterns from Mobile Sensing Data: A Case Study on Blood Glucose Inference

**Miao He**
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University

**Weixi Gu**[*]
guweixigavin@gmail.com
University of California, Berkeley
China Academy of Industrial Internet

**Ying Kong**
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University

**Lin Zhang**
Tsinghua-Berkeley Shenzhen Institute
Tsinghua University

## ABSTRACT

The high-dimensional and co-evolved data streams sensed by mobile devices typically exists time delays that form the "causal-and-effect" patterns. Understanding the informative causal patterns from the multivariate time series is critical but challenging for the inference tasks with the sensing data. While a large scope of statistical learning methods has undergone great advances in the causal pattern recognition problem, most of them are still limited in the unreliable causal analysis, high computational complexity and the environmental noise interruption. To this end, we propose a novel directed information (DI)-aided approach to efficiently select the casual patterns from a set of feature streams collected from mobile devices. The proposed approach has been evaluated on a real blood glucose sensing dataset. The results demonstrate our proposed approach outperforms the traditional methods in cost efficiency and inference accuracy.

## CCS CONCEPTS

• **Applied computing** → **Health care information systems**; • **Information systems** → *Information systems applications.*

## KEYWORDS

Causal pattern mining; Multi-variate sensing data stream; Blood glucose inference

[*] Corresponding Author

*UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom*
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6869-8/19/09.
https://doi.org/10.1145/3341162.3343820

**ACM Reference Format:**
Miao He, Weixi Gu, Ying Kong, and Lin Zhang. 2019. Poster: Identifying Causal Patterns from Mobile Sensing Data: A Case Study on Blood Glucose Inference. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 9–13, 2019, London, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3341162.3343820

## 1 INTRODUCTION

With the recent advancements in instrumentation and measurement technologies, researchers in various disciplines now have access to rich, real-time sensing data [1, 4, 5]. Among the vast quantity of information generated by such processes, some features are highly correlated with the target application, while others may be less relevant or even redundant. For many data mining tasks, using input that contains irrelevant or redundant patterns will not help but rather hurt their performance. This further lead to a problem of subset selection for pattern recognition. Consequently, the goal of pattern mining can be summarized as identifying the most informative causal patterns from the observed data, with a given target.

The causal patterns typically occur in time series [2, 3]. Figure 1 illustrates four typical patterns which we meet in processing the sensing data. The orange shadow area denotes the occurrence of target event. As shown, pattern 1 and pattern 2 are irrelevant patterns, as they do not provide any useful information about the interested target. Pattern 3 is a causal pattern, while pattern 4 is a relevant but unpredictable pattern, as it only begun to change after the target event already happened.

Since we want to implement our recognized patterns in inference tasks, causal pattern discovery becomes necessary because it can identify the precedents of target while normal

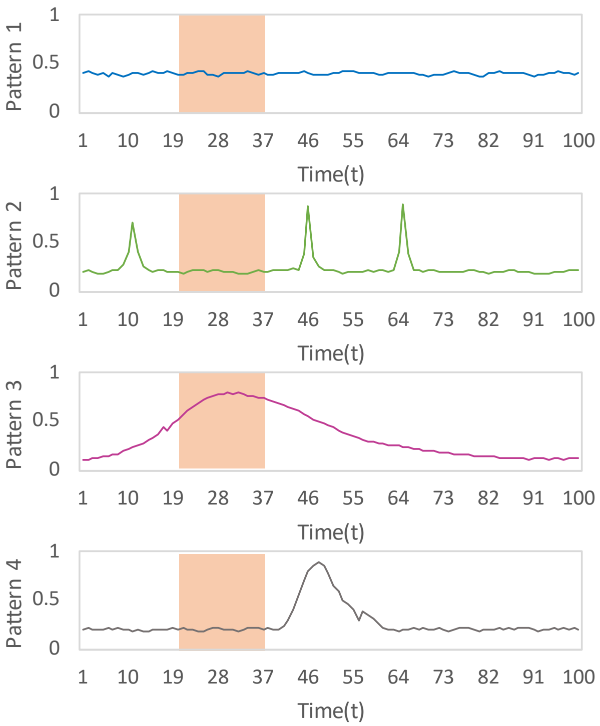


**Figure 1: Illustration of four typical patterns in sensing data.**

pattern mining cannot. There are two kinds of causality definitions in literature. One regards to time series prediction(eg. Granger Causality) while the other one is about counterfactuals [9]. In this paper,we focus on the first definition, which works in statistical meaning with the involvement of temporality. It is suitable when applying to sensing data, considering the sensing streams are mostly multi-dimensional series evolving over time, the rich temporal information embodied in the sensing data can be exploited in causal analysis.

In this work, we describe the characteristics of sensing data from two aspects: multi-dimensionality and temporality. Multi-dimensionality incurs a problem of selection, as not all the series are useful in predicting/reporting our interested targets or events. While temporality plays a critical role in causal relationship discovery, and further facilitate the later inference task.

Targeted on these characteristics, we put forward an innovative causal pattern mining framework for inference tasks. With the proposed method, the following aspects would benefit:

(1) The computation cost of an inference task can increase dramatically with the growing of potentially related features. With the picked out comparatively rather small causal feature set, the computation cost can be decreased greatly.

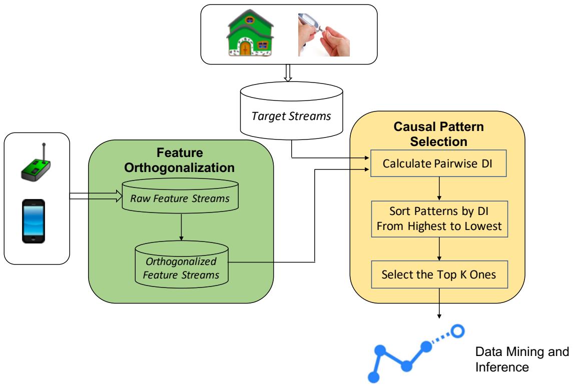

Target Streams
Feature Orthogonalization
Raw Feature Streams
Orthogonalized Feature Streams
Causal Pattern Selection
Calculate Pairwise DI
Sort Patterns by DI From Highest to Lowest
Select the Top K Ones
Data Mining and Inference

**Figure 2: The framework of our proposed method.**

(2) With leveraging causal pattern mining, large amounts of noisy information can be filtered out, so as to enhance the accuracy of the inference task in a later stage.

(3) In this work, we solve the cardinality-constraint NP-hard selection problem in two phases. The adoption of feature orthogonalization helps suppress the correlation among feature streams, then the DI-aided pattern mining method selects causal patterns with linear complexity.

(4) Prior studies [7, 10] in blood glucose inference usually ignored the causal analysis. Our work filled in this gap and is a useful supplement.

## 2 FRAMEWORK FOR CAUSAL PATTERN MINING

There are two modules in the proposed framework: feature orthogonalization and causal pattern selection. As shown in Figure 2.

### Feature Orthogonalization

Usually, the collected sensing data are correlated and in order to get an efficient solution with a cardinality constraint, we should prevent the inclusion of duplicated information. This is the motivation for this module. Moreover, as we want to provide better interpretability, we choose symmetric orthogonalization, which offers the highest similarity between the transformed vectors and their corresponding original vectors among existing orthogonalization methods. Through a series of transformations, a set of orthogonalized features are obtained with removed correlation. Thus, we can select the feature series one by one, without worrying about the duplicated information. Furthermore, as symmetric orthogonalization disturbs the original features the least compared with other methods, the transformed features corresponds very well with the original ones.

**Table 1: Blood Glucose Inference Accuracy using HMM and CRF.**

| | K=5 | K=10 | K=54(all patterns) |
|---|---|---|---|
| **HMM** | 64.1% | **64.2%** | 58.4% |
| **CRF** | 69.1% | **69.4%** | 68.0% |

### Causal Pattern Selection

In this module, we adopt a causal metric, DI, to measure the causal information from feature series to the target series. As a tool for causal analysis, DI is not only a qualitative metric, but also a quantitative metric. As it can provide a numeric result for causal strength comparison besides identifying the existence of causality. In this module, with the orthogonalized feature streams, we can calculate the directed information between feature and target stream one by one, so as to subtly avoid the difficulty of estimating high-dimensional directed information. The causal pattern mining problem is transformed into a directed information maximization problem. Such design reduces the computation complexity greatly.

The combination of the two modules helps identify the informative causal patterns effectively. First, through orthgonalizing the feature vectors by symmetric method, the correlation within the features are removed while the correspondence and similarity are mostly preserved. This further simplifies the subset selection problem. Second, with the aid of the quantitative causal metric DI, the orthogonalized features can be selected with linear complexity, noises and non-causal features are removed so as the most contributive causal ones are remained to enhance the inference accuracy and efficiency.

## 3 EXPERIMENTS AND EVALUATION

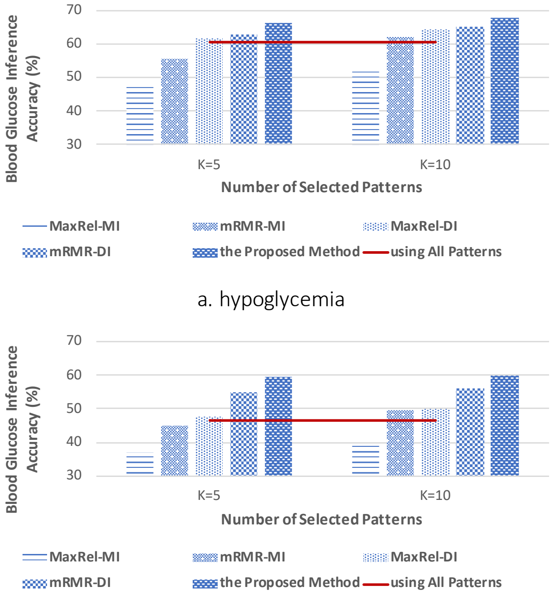
### Dataset Introduction.

Accurate blood glucose prediction is beneficial for human health. Many previous works have selected features empirically, which has little confidence on the relevancy or the redundancy among the selected features, resulting in a disappointing performance. In one recent work [6] 54 features were already extracted for blood glucose level estimation. Such high dimensional features are suitable for deep neural network training but may not work with simpler models such as HMM and CRF. Therefore, we searched for a subset of causal patterns that would also guarantee the performance of these simple inference models.

We evaluate our causal pattern mining method on a blood glucose inference dataset in [6]. The dataset is composed of two kinds of features: physiological and temporal features. Physiological features describe carbohydrate and insulin dynamics, as well as their fluctuations influenced by daily sleep, activities, food and drug intakes. The temporal features involve the average blood glucose concentration over the past 5 days, and physiological factors over the past 5 minutes.

**Table 2: List of comparing methods.**

| Criterion | Expression |
|---|---|
| MaxRel-MI | $\max I_r(S_x, y) = \frac{1}{\lvert S_x \rvert} \sum_{x_i \in S_x} I(x_i; y)$ |
| mRMR-MI | $\max I_{mr}(I_r, y) = \frac{1}{\lvert S_x \rvert} \sum_{x_i \in S_x} I(x_i; y) - \frac{1}{\lvert S_x \rvert^2} \sum_{x_i, x_j \in S_x} I(x_i; x_j)$ |
| MaxRel-DI | $\max I_r(S_x \rightarrow y) = \frac{1}{\lvert S_x \rvert} \sum_{x_i \in S_x} I(x_i \rightarrow y)$ |
| mRMR-DI | $\max I_{dr}(I_r, y) = \frac{1}{\lvert S_x \rvert} \sum_{x_i \in S_x} I(x_i \rightarrow y) - \frac{1}{\lvert S_x \rvert^2} \sum_{x_i, x_j \in S_x} I(x_i \rightarrow x_j)$ |



a. hypoglycemia

b. hyperglycemia

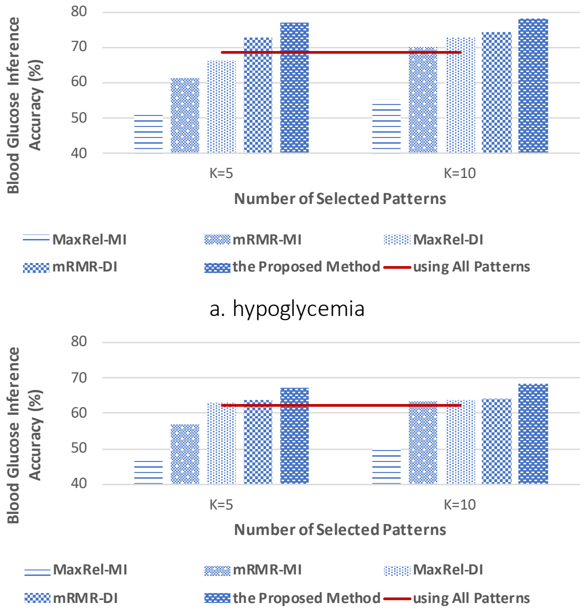
**Figure 3: Performance of blood glucose inference using HMM with a subset of 5 and 10 causal patterns and the entire feature set.**

### Selected Patterns Using the Proposed Method.

Specifically, among the total 54 potential patterns, the carbohydrate and insulin dynamics, and the average blood glucose concentration over the past 5 minutes are the three most contributive patterns. Such results also coincide with medical results [8], blood glucose levels are causally determined by the carbohydrate and insulin levels, and it has lagging influence caused by the previous blood glucose level.

### Classification Performance.

Firstly, we summarize the overall accuracy in Table 1 using our proposed method. As shown, the best overall accuracy



a. hypoglycemia

b. hyperglycemia

**Figure 4: Performance of blood glucose inference using CRF with a subset of 5 and 10 causal patterns and the entire feature set.**

was achieved using the subset of causal patterns ($K = 10$) selected by our causal feature selection method, and it is slightly superior to the 5-pattern subset ($K = 5$). However, from the perspective of efficiency, $K = 5$ is also a good option as it use half the size of patterns, while the performance does not hurt much. What is surprising is that, the accuracy of the entire feature set (i.e. $K = 54$) is lowest. The involvement of too much noise may lead to the deterioration of inference performance.

More specifically, Figure 3 and Figure 4 illustrate the performance of blood glucose inference of the proposed method and other four methods shown in Table 2, with HMM and CRF, respectively. As we can see, the proposed method outperfoms other four methods significantly with either HMM or CRF, under both hypoglycemia inference and hyperglycemia inference.

Finally, the inference accuracy when using a subset is superior to the performance of using $K = 54$ whole features, with both HMM and CRF. This result demonstrates that the optimal causal pattern subset selection enables the prediction models to avoid overfitting and improves their generalization ability. Useful and predictive causal patterns are picked, and lots of computation power is saved in every prediction as the selected subset is usually rather small compare to the original set.

## 4 CONCLUSIONS

Causal pattern mining is a curcial problem in multivariate sensing data analysis. In this work, we have designed a generic two-module causal pattern selection framework on multivariate sensing data mining. In the first phase, feature streams got orthogonalized with the mildest disturbance and steady corresponding relationships, and this paved the way for module two. A causal pattern selection method with linear complexity was introduced in this latter module, which highly enhances the efficiency for pattern selection compared with other existing methods. Finally, extensive experiments were conducted with a real world dataset, demonstrate the effectiveness of out proposed method.

## REFERENCES

[1] Adnan Akbar, Abdullah Khan, Francois Carrez, and Klaus Moessner. 2017. Predictive analytics for complex IoT data streams. *IEEE Internet of Things Journal* 4, 5 (2017), 1571–1582.

[2] Qianjin Du, Weixi Gu, Lin Zhang, and Shao-Lun Huang. 2018. Attention-based LSTM-CNNs For Time-series Classification. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 410–411.

[3] Weixi Gu, Yunxin Liu, Yuxun Zhou, Zimu Zhou, Costas J Spanos, and Lin Zhang. 2017. BikeSafe: bicycle behavior monitoring via smartphones. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 45–48.

[4] Weixi Gu, Longfei Shangguan, Zheng Yang, and Yunhao Liu. 2015. Sleep hunter: Towards fine grained sleep stage tracking with smartphones. *IEEE Transactions on Mobile Computing* 15, 6 (2015), 1514–1527.

[5] Weixi Gu, Zheng Yang, Longfei Shangguan, Xiaoyu Ji, and Yiyang Zhao. 2014. Toauth: Towards automatic near field authentication for smartphones. In *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE, 229–236.

[6] Weixi Gu, Yuxun Zhou, Zimu Zhou, Xi Liu, Han Zou, Pei Zhang, Costas J Spanos, and Lin Zhang. 2017. SugarMate: Non-intrusive Blood Glucose Monitoring with Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 54.

[7] Weixi Gu, Zimu Zhou, Yuxun Zhou, Miao He, Han Zou, and Lin Zhang. 2017. Predicting Blood Glucose Dynamics with Multi-time-series Deep Learning. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 55.

[8] Boris P Kovatchev, William L Clarke, Marc Breton, Kenneth Brayman, and Anthony McCall. 2005. Quantifying temporal glucose variability in diabetes via continuous glucose monitoring: mathematical methods and clinical application. *Diabetes technology & therapeutics* 7, 6 (2005), 849–862.

[9] Judea Pearl. 2003. Causality: models, reasoning, and inference. *Econometric Theory* 19, 675-685 (2003), 46.

[10] Kevin Plis, Razvan Bunescu, Cindy Marling, Jay Shubrook, and Frank Schwartz. 2014. A machine learning approach to predicting blood glucose levels for diabetes management. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.