Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240807

Surveillance of Disease Outbreaks Using Unsupervised Uni-Multivariate Anomaly Detection of Time-Series Symptoms

Atiye Sadat HASHEMI[a,b,1,2], Mirfarid Musavian GHAZANI[a,2], Mattias OHLSSON[a,c], Jonas BJÖRK[b,d] and Dominik DIETLER[b]
[a] *Center for Applied Intelligent Systems Research, Halmstad University, Sweden*
[b] *Division of Occupational and Environmental Medicine, Lund University, Sweden*
[c] *Centre for Environmental and Climate Science, Lund University, Sweden*
[d] *Clinical Studies Sweden, Forum South, Skåne University Hospital, Lund, Sweden*
ORCiD ID: Atiye Sadat Hashemi https://orcid.org/0000-0001-5191-0424

Abstract. Effectively identifying deviations in real-world medical time-series data is a critical endeavor, essential for early surveillance of disease outbreaks. This paper demonstrates the integration of time-series anomaly detection techniques to develop surveillance systems for disease outbreaks. Utilizing data from Sweden's telephone counseling service (1177), we first illustrate the trends in physical and mental symptoms recorded as contact reasons, offering valuable insights for outbreak detection. Subsequently, an advanced anomaly detection technique is applied incrementally to these time-series symptoms as univariate and multivariate approaches to assess the effectiveness of a machine learning-based method on early detection of the COVID-19 outbreak.

Keywords. Anomaly detection, Anomaly transformer, COVID-19 pandemic, Incremental learning, Public health surveillance.

1. Introduction

Recent advances in technology allow for collecting large amounts of data over time in healthcare [1]. This capability is crucial for effective surveillance of disease outbreaks, as it enables the detection of subtle trends and patterns that may indicate emerging threats. Longitudinal data allows for the identification of seasonal variations, geographic hotspots, and demographic disparities in disease prevalence, aiding in targeted intervention strategies. In addition, harnessing the power of cutting-edge artificial intelligence, further enhances the ability to analyze large datasets efficiently, leading to more timely and accurate disease surveillance [2]. When considering disease outbreaks as an event of interest, anomaly detection stands as a pivotal method.

Anomaly detection, by enabling the timely identification of unusual patterns or deviations from expected norms in epidemiological data, plays a significant role in the surveillance of disease outbreaks [3]. Leveraging various statistical and machine

[1] Corresponding Author: Atiye Sadat Hashemi; E-mail: atiye_sadat.hashemi@med.lu.se.
[2] Authors contributed equally.

learning techniques, anomaly detection algorithms sift through large volumes of data to identify abnormal occurrences indicative of potential disease outbreaks [4]. These anomalies could manifest as unexpected spikes in reported cases, unusual geographic distributions, or atypical symptom profiles. By detecting these deviations early on, public health authorities can swiftly initiate targeted interventions, such as increased surveillance, resource allocation, or public health messaging, to mitigate the spread of infectious diseases and minimize their impact on public health [5].

Unsupervised anomaly detection techniques are extensively explored as a practical solution to real-world problems because of the unpredictable nature of anomalous events. Unsupervised time series anomaly detection can be classified into two main categories known as univariate and multivariate approaches, each of which can be point-wise or subsequence-wise [6]. Numerous methodologies are available for time series anomaly detection, ranging from traditional statistical methods to sophisticated machine learning algorithms such as isolation forests [7], autoencoders [8], copula-based outlier detectors [9], graph neural networks (GNNs) [10], generative adversarial networks (GANs) [11], and anomaly transformers [12], etc.

In the context of disease outbreaks, it is paramount to address both univariate and multivariate scenarios; while anomalies might not be evident in individual symptoms within a univariate signal, the overall multivariate signal might exhibit abnormalities, or vice versa. Although an effective multivariate anomaly detection model should ideally include separate univariate analyses, different models have different biases and it has been shown that these models can fail to capture these anomalies in univariate subspaces [13]. In this paper, as preliminary research, we emphasize the importance of monitoring uni-multivariate signals to ensure a comprehensive understanding and detection of disease outbreaks. The rest of the paper is organized as follows. Sections 2 and 3 describe the framework proposed for surveillance of disease outbreaks and the experiments, respectively. In Section 4, we conclude the paper by presenting our future work.

2. Methodology

We propose an incremental learning approach aimed at detecting anomalies within time series data, with a specific focus on the early identification of disease outbreaks. The process involves continuously updating the understanding of the model of normal patterns as new data arrives, enabling real-time detection of deviations from expected behavior. This approach allows the system to adapt dynamically to evolving trends and anomalies. Figure 1 illustrates the proposed framework, which leverages both univariate and multivariate anomaly detection models in combination.

As shown in Figure 1, in a scenario of incremental learning, the anomaly detection model (Anomaly Transformer [12]) is initially trained using a subset of data (two months), allowing it to grasp the normal and anomaly patterns within the temporal data. Subsequently, the model is iteratively retrained and tested using additional data (spanning one week) which is incorporated into the analysis. This iterative process continues until the entire dataset has been utilized. The overarching objective is to enable the timely detection of disease outbreaks, leveraging the gradual accumulation of data to enhance the sensitivity of the model to emerging anomalies within the temporal dynamics of the observed phenomena. This methodological framework underscores a systematic and data-driven approach towards proactive disease

surveillance and early intervention strategies. In this section, the dataset, mathematical notation, and time series anomaly detection model are discussed.

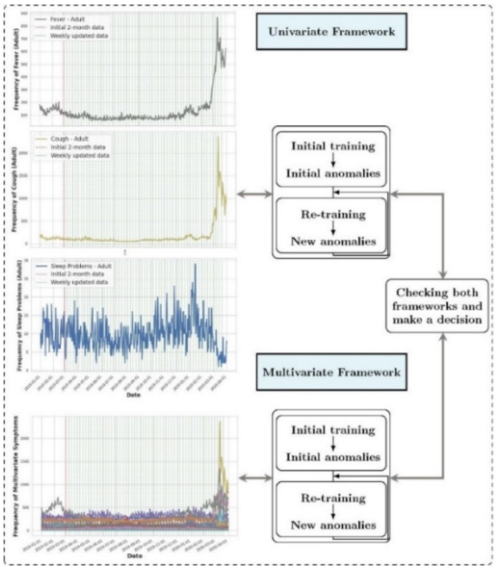

Figure 1. The proposed framework for the surveillance of disease outbreaks.

2.1. Dataset

The data is the calling data to the 1177 hotline, the Swedish Healthcare Guide phone line which provides medical advice on care and illnesses[3]. The variable *Contact Reason* is the main variable of interest in this study, which includes 190 symptoms ranging from common issues such as cough (adult/child), fever (adult/child), bloody cough, chest pain, abdominal pain (adult/child), chest pain, sleep problems. Currently, these symptoms are manually recorded by a healthcare professional. The dataset spans from January 1st, 2019, to June 3rd, 2021, encompassing the period of heightened activity related to the coronavirus pandemic.

2.2. Mathematical notation and time-series anomaly detection model

The mathematical notation of incremental anomaly detection in univariate and multivariate scenarios in our framework can be described as follows.

Univariate Framework: Let $X_i(t)$ denote the temporal signal for symptom i at time t where $i = \{1,2, \dots, m\}$ and $t = \{1,2, \dots, T\}$. Let $A_i(t)$ represent if symptom i at time t is an anomaly or not. In initialization step, for the initial dataset $D_0(Time, S_1, S_2, \dots, S_m)$, $A_i(t) = M\big(X_i(t)\big)$ where $(t, i) \in D_0[Time, S_i]$. The function M represents the anomaly detection algorithm, which takes a dataset as input and returns whether each data point is an anomaly or not. For each subsequent time window

[3] For further details, please refer to the Inera website (https://www.inera.se/tjanster/1177/).

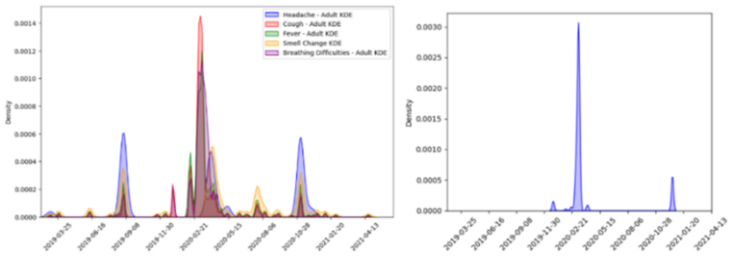

Figure 2. KDE plots illustrating the concatenated anomaly estimation in the incrementally learning using our framework on 1177 symptoms dataset. Left: Five randomly selected anomaly estimation results using univariate symptoms, and Right: anomaly estimation results using multivariate framework.

$w_t(Time, S_1, S_2, \dots, S_m)$, $D_t = D_{t-1} \cup w_t$ and $A_i(t) = M\big(X_i(t)\big)$ where $(t, i) \in D_t[Time, S_i]$.

Multivariate Framework: Let $X(t)$ denote the multivariate temporal signal at time t where $X(t) = \{X_i(t)\}$ and $i = \{1,2, \dots, m\}$. Let $A(t)$ show that if there is an anomaly at time t or not. In initialization step, for the initial dataset $D_0(Time, S_1, S_2, \dots, S_m)$, $A(t) = M\big(X(t)\big)$ where $(t, i) \in D_0[Time, S_i]$. Same as before, the function M represents the anomaly detection algorithm. For each subsequent time window $w_t(Time, S_1, S_2, \dots, S_m)$,$D_t = D_{t-1} \cup w_t$ and $A(t) = M\big(X(t)\big)$ where $t \in D_t[Time]$.

We implemented our framework utilizing one of the state-of-the-art methods known as Anomaly Transformers [12]. Anomaly Transformer has adapted the transformers to perform unsupervised anomaly detection in the time series. The model uses the self-attention map to detect temporal associations, and an adjacent-concentration prior to take into account the rarity of anomalies as well as the fact that adjacent time points share similar abnormal patterns. In our framework, the time series anomaly detection mentioned as M is an Anomaly Transformer. In the model configuration, a sequence length of 24 hours and a stride of 1 are employed, along with three anomaly transformer blocks each featuring eight heads, while the dataset for training is estimated to have an anomaly ratio of 0.001.

3. Results

The kernel density estimate (KDE) plots of estimated anomalies illustrating the outcomes derived from our dataset, encompassing both univariate and multivariate scenarios within our incremental learning-based framework (refer to Figure 1), are presented in Figure 2. All the steps are concatenated to be shown in one plot. These results demonstrate how anomalies in the frequency of particular symptoms (five randomly selected) and their combined analysis in the multivariate context can effectively detect the emergence of significant COVID-19 peaks. Our subsequent endeavor involves leveraging domain expertise to interpret additional abnormalities.

4. Conclusion and Future Work

Anomaly detection frameworks can adapt dynamically to evolving epidemiological landscapes, enhancing their effectiveness in detecting emerging threats and supporting proactive decision-making in disease surveillance and control efforts. In this paper, we propose a uni-multivariate anomaly detection framework based on time series signals of symptoms. The framework aims to update the model by training it on the incoming observations, computing scores, and estimating anomalies. In future work, we will focus on incorporating spatiotemporal signals [14] into the framework to consider subgroups defined by space, time, and population characteristics in disease surveillance. This expansion aims to enhance the accuracy of the surveillance efforts by integrating geographic and demographic dimensions to detect and respond to emerging health threats more effectively. We will also leverage expert knowledge to interpret the results of the models to identify both seasonal and non-seasonal anomalies. By combining advanced modeling techniques with expert insights, we aim to develop systems to address emerging health threats promptly and effectively.

Acknowledgment and Ethical considerations: This study is part of the SWECOV project. Ethical permission is granted by the Swedish Ethical Review Authority (permit numbers 2020-06492, 2021-01115, 2022-01355-02, and 2022-06118-02). This work was supported by grants from the Lars Mikael Karlsson Foundation (LMK-stiftelsen) and from the Swedish Research Council (VR; dnr 2022-06358).

References

[1] Lundström J, Hashemi AS, Tiwari P. Explainable Graph Neural Networks for Atherosclerotic Cardiovascular Disease. In33rd MIE2023, Gothenburg, 22-25 May 2023, Code 189285 2023 May 1 (Vol. 302, pp. 603-604). IOS Press.
[2] Xu Z, Su C, Xiao Y, Wang F. Artificial intelligence for COVID-19: battling the pandemic with computational intelligence. Intelligent medicine. 2022 Feb 1;2(1):13-29.
[3] Karadayi Y, et al. Unsupervised anomaly detection in multivariate spatio-temporal data using deep learning: early detection of COVID-19 outbreak in Italy. Ieee Access. 2020 Sep 7;8:164155-77.
[4] Ali O, Ishak MK, Bhatti MK. Early COVID-19 symptoms identification using hybrid unsupervised machine learning techniques. Computers, Materials and Continua. 2021 Jan 1;69(1):747-66.
[5] Homayouni H, Ray I, Ghosh S, Gondalia S, Kahn MG. Anomaly detection in COVID-19 time-series data. SN Computer Science. 2021 Jul;2(4):279..
[6] Blázquez-García A, Conde A, Mori U, Lozano JA. A review on outlier/anomaly detection in time series data. ACM Computing Surveys (CSUR). 2021 Apr 17;54(3):1-33.
[7] Karczmarek P, Kiersztyn A, Pedrycz W, Al E. K-means-based isolation forest. Knowledge-based systems. 2020 May 11;195:105659.
[8] Khoshbakhtian F, Ashraf AB and Khan SS. Covidomaly: A deep convolutional autoencoder approach for detecting early cases of covid-19. 2020;arXiv preprint arXiv:2010.02814.
[9] Li Z, Zhao Y, Botta N, Ionescu C and Hu X. COPOD: copula-based outlier detection. In 2020 IEEE international conference on data mining (ICDM) (2020) (pp. 1118-1123). IEEE.
[10] Deng A and Hooi B. Graph neural network-based anomaly detection in multivariate time series. In Proceedings of the AAAI conference on artificial intelligence 2021;35(5): 4027-4035.
[11] Miao J, Tao H, et al. Reconstruction-based anomaly detection for multivariate time series using contrastive generative adversarial networks. Information Processing & Management. 2024;61,103569.
[12] Xu J, Wu H, Wang J and Long M. Anomaly transformer: Time series anomaly detection with association discrepancy. 2021 arXiv preprint arXiv:2110.02642.
[13] Aggarwal CC, 2016. Outlier analysis second edition.
[14] Karadayi Y, et al. Unsupervised anomaly detection in multivariate spatio-temporal data using deep learning: early detection of COVID-19 outbreak in Italy. 2020, Ieee Access, 8, pp.164155-164177.