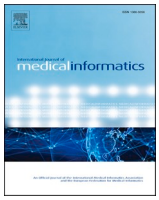
International Journal of Medical Informatics 194 (2025) 105696

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Smart data-driven medical decisions through collective and individual anomaly detection in healthcare time series

Farbod Khanizadeh [a], Alireza Ettefaghian [b], George Wilson [c], Amirali Shirazibeheshti [d], Tarek Radwan [d], Cristina Luca [e,*]

[a] Operation & Information Management, Aston Business School, Birmingham
[b] Anglia Ruskin University, Cambridge
[c] School of Computing and Information Science, Anglia Ruskin University, Cambridge
[d] AT Medics, London
[e] School of Computing and Information Science, Anglia Ruskin University, Cambridge CB1 1PT, United Kingdom

ARTICLE INFO

Keywords:
Decision making
Health forecasting
Outliers
Time series
Anomaly Detection
Unsupervised learning

ABSTRACT

Background: Anomalies in healthcare refer to deviation from the norm of unusual or unexpected patterns or activities related to patients, diseases or medical centres. Detecting these anomalies is crucial for timely interventions and efficient decision-making, helping to identify issues like operational inefficiencies, fraud and emerging health complications.

Objectives: This study presents a novel method for detecting both collective and individual anomalies in healthcare data through time series analysis using unsupervised machine learning. The dual-strategy approach leverages two methodologies: a 'practice centre-based approach' which monitors changes across different practice centres and a 'process-based approach' which focuses on identifying anomalies within individual centres. The former allows for early detection of systemic issues, while the latter highlights specific irregularities within a centre's operations.

Methods: The study utilised a dataset over 500,000 medical records from multiple GP practice centres in the UK collected between 2018–2023. Data are clustered using DBSCAN to identify collective anomalies from deviations from linear trends in consecutive two-month scatterplots. Individual anomalies are identified by examining the SOM-clustered time series of various medical processes within a specific practice centre, where graphs show deviation from the typical pattern.

Findings: Our approach addresses some challenges posed by the complexity and sensitivity of healthcare data by not requiring personal information. The method offers accurate visual representations making the data accessible and interpretable for non-technical users. Unlike traditional methods focusing solely on subsequence anomalies, our technique analyses the collective behaviour across multiple time series providing a more comprehensive perspective.

Conclusion: This study underscores the importance of integrating unsupervised anomaly detection with clinical expertise to ensure that statistically anomalous patterns align with clinical relevance. The dual-strategy clustering method holds significant potential for enabling timely interventions, proactively identifying potential crises, and ultimately contributing to better decision-making and operational efficiency within the healthcare sector.

1. Introduction

Anomalies in healthcare involve unusual conditions or activities by patients, diseases, or medical centres Samariya et al. [47]). Detecting these anomalies provides early warnings, enabling timely interventions and efficient decision-making. However, what the system flags as anomalies may not always match clinical expectations. Anomaly detection in healthcare is crucial for identifying issues like operational

* Corresponding author.
E-mail addresses: khanizaf@aston.ac.uk (F. Khanizadeh), ae36@aru.ac.uk (A. Ettefaghian), George.Wilson@aru.ac.uk (G. Wilson), a.shirazibeheshti@nhs.net (A. Shirazibeheshti), tradwan@nhs.net (T. Radwan), cristina.luca@aru.ac.uk (C. Luca).

https://doi.org/10.1016/j.ijmedinf.2024.105696
Received 20 June 2024; Received in revised form 6 November 2024; Accepted 7 November 2024
Available online 17 November 2024

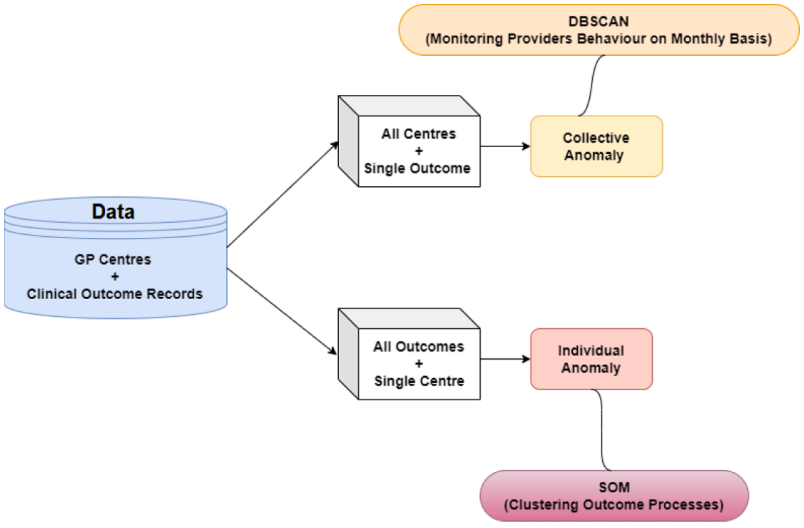
Table 1
Key Themes and Literature on Anomaly Detection Methods.

Theme	Research Sources
Anomaly and its Types	Early work addressing anomaly detection is described by Grubbs [16] who defines an outlier as an observation that appears to deviate markedly from other members of the sample in which it occurs. According to Chandola et. al. [8] they can be contextual (anomalous but conditional on other parameters), collective (a subset of data anomalous with respect to the entire dataset) or global (a single instance of data anomalous if it's too far off from the rest). Foorthuis [13] further extends this understanding with a comprehensive typology of anomaly types and subtypes based on fundamental dimensions of anomalies, providing a structured framework for classifying these deviations in data.
Anomaly Detection Techniques and Methods	Various techniques have been reported to detect anomalies in datasets each with their advantages and disadvantages, including approaches that are classification-based Bergman & Hoshen[3],Steinwart [55],Wei [66], clustering-basedDuan et. al. [9],Pamula et. al. [40],Samriya et. al. [46],Kayhan et. al. [19], information theoretic Wu & Wang[67],Lee & Xiang [27], statistical Goldstein & Uchida[14]and spectralShyu et. al. [52],Erdogan [11]. Thudumu et. al. [61] provides an in-depth analysis of anomaly detection techniques particularly in the context of high-dimensional big data, where computational complexity can increase significantly highlighting the need for efficient algorithms to ensure the processing time doesn't become prohibitively long. Ngai et al. [38] presents the first systematic review of data mining techniques for financial fraud detection, analyzing 49 journal articles from 1997 to 2008 and categorizing them by fraud type and technique. Similarly, Pourhabibi et al. [41] develops a framework to synthesize literature on graph-based anomaly detection (GBAD) methods for fraud detection from 2007 to 2018, investigating trends, identifying challenges, and providing recommendations to enhance credibility.
Time Series Anomaly Detection Techniques in Healthcare	The detection of anomalies in healthcare data using time series analyses by unsupervised learning is a fundamental strategy according to Li et. al. [28]. Classical techniques for time series outlier detection include the traditional moving average approaches (ARMA, ARIMA) or training artificial neural networks (ANN). As the ANNs require large datasets, researchers have identified other approaches. For example, Sun et. al. [57] presents an optimised unsupervised outlier detection method using Probabilistic Suffix Trees (PSTs), achieving effectiveness with a significantly smaller PST (less than 5 % of the original size) in protein sequence experiments. Benkabou et. al. [2] introduce a weighted clustering approach based on entropy and dynamic time warping, specifically tailored for time series data. Outliers are identified through the optimization of a novel cost function designed for this type of data. Tsay et. al. [62] and Blázquez-García et. al. (2021) present methods which extend the concept of outliers from univariate to multivariate time series analysis, including identification of multivariate anomalies (e.g. joint and marginal test statistics) and algorithm optimisation (e.g. dimensionality reduction, shapelet-based learning). Keogh et. al. [21] focuses on operational 'surprise' using Markov models encoded efficiently using a suffix tree. Their approach is capable of processing large datasets in linear time and space, making it a practical tool for discovering anomalous patterns in time series without prior specification of what constitutes an anomaly. Wang et. al. [65] presents an anomaly detection method for time series data streams, addressing the challenges of high dimensionality, large data volumes and rapid updates characteristic of time series data by using multiple random convolution kernels for feature transformation. As the approach does not require pre-trained models, it is efficient with potential for further improvement for more demanding applications.
Clustering Approaches for Time-Series Analysis	Two notable clustering approaches used by analysts of time-series data are the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and the Self-Organizing Map (SOM). DBSCAN is known for its ability to discover clusters of arbitrary shapes and sizesÇelik et. al. [6],Sheridan [49],Jain et. al. [18]. Furthermore, it efficiently identifies outliers or noisy data points and does not require the user to specify the number of clusters beforehand. This is beneficial in situations where the optimal number of clusters is not known in advance or when the data may contain varying numbers of clusters Khan et. al. [22],Monalisa & Kurnia [35]. SOM on the other hand preserves the topological properties of the input data Brereton[4],Kohonen [25],Nowak-Brzezińska & Horyń [39]. Massi et. al. [32] applied an unsupervised clustering method for outlier detection within administrative healthcare databases. Shirazibeheshti et. al. [51] employed mean shift clustering to cluster risk vectors (medication risk scores) into different risk groups and identify any potential anomality in prescription records.
Comparative Analysis of Healthcare Anomaly Detection Methods	Different approaches have been considered by various workers[45,43,53,12]. Ikono et. al. [17] reviewed anomaly detection methods in healthcare based on a study of 88 works in the academic and professional published literature between 1984 and 2016. Their work concluded that fuzzy logic excels in handling uncertainty but lacks precision, neural networks are adept at pattern recognition but can be computationally intensive Brockett et. al. [5], Bayesian networks effectively incorporate probabilistic relationships but struggle with large datasets Ekina et. al. [10], and clustering analysis is useful for identifying patterns but is sensitive to outliers Liu, Q. and Vasarhelyi [30].
Healthcare Fraud Detection and Prevention	Healthcare fraud is a major issue globally and anomaly detection is a principal means of combating this issue [33,68,54,1,63,58,50,31]. For example, Thornton et. al. [60] addresses the pervasive issue of healthcare fraud in the U.S. Medicaid system (estimated to cost around $700 billion annually). Their use of unsupervised data mining techniques (including outlier detection) led to 71 % of the top suspicious providers being referred for investigation. Their specific method was primarily concerned with dental provision and this speciality-specific approach may limit its immediate applicability to other medical domains without further development. Another example is the introduction of a data-driven system called UNISIM designed to detect fraudulent Medicare claimants in Australia Tang [59]. Despite the complexity of accurately defining irregularity, the system autonomously identifies patterns of prescription abuse over periods of 1 to 4 years. Other work in a UK context highlights similar concerns regarding the level of healthcare fraud and discusses appropriate data mining techniques Kirlidog & Asuk, [23]. Recent advancements by Kuo & Pham, [26] explore the use of learning models to identify misconduct without compromising patient data privacy, making it a promising approach for broad application across various healthcare settings.

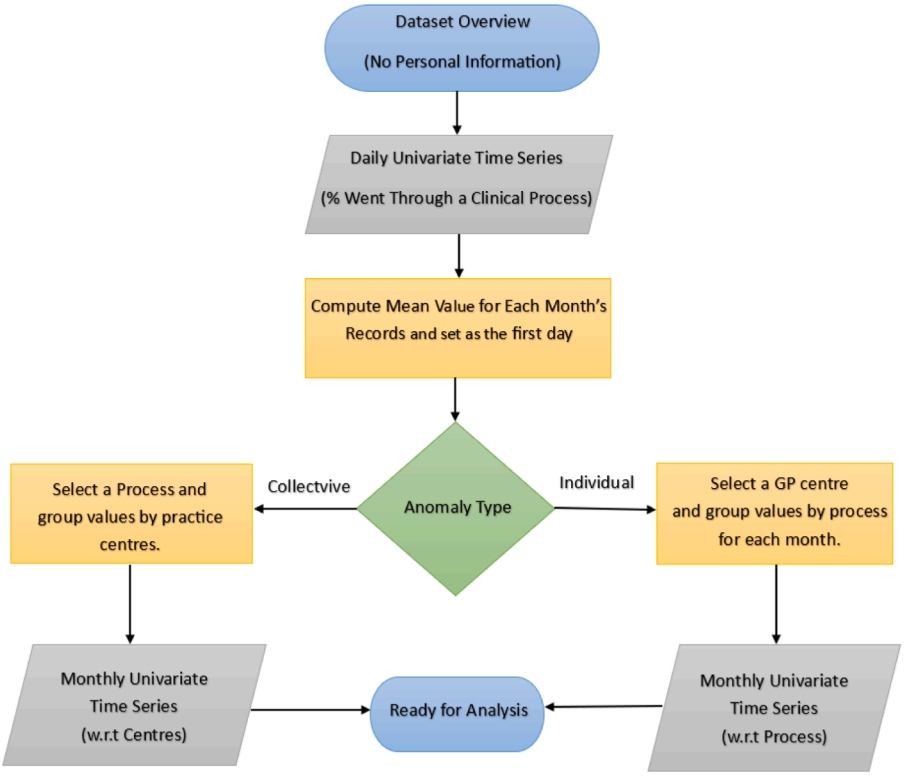
inefficiencies, fraud, and emerging health complications Stylianou et al. [56],Griffith [15]. The main goal is to identify patterns that deviate from norms. This task is challenging due to the complexity and sensitivity of healthcare data[36]. Failure to detect anomalies can result in financial losses, compromised patient care, and reputational damage.

While numerous anomaly detection methods have been developed, each comes with its own set of limitations (Table 1). Traditional statistical methods often struggle with the high-dimensional and complex nature of healthcare data, leading to either false positives or missed anomalies. Machine learning approaches, such as neural networks, can be computationally intensive and require large datasets for training which may not always be available. Furthermore, many existing methods focus on point or subsequence anomalies within a single time series, often overlooking collective anomalies that appear across multiple related series. Additionally, the need for labelled data in supervised learning models poses a significant challenge as such data are rare and difficult to obtain in healthcare.

An individual anomaly refers to a single time series exhibiting irregular patterns compared to others. On the other hand, a collective anomaly arises from a group of related instances, each of which may not appear anomalous in isolation but collectively deviates from the norm. A collective anomaly within a time series is where a sequence of data points exhibits an unusual pattern. Deviations of several practice centres from the norm are considered, where different data points from different



(a)



(b)

Fig. 1. Process flow Overview for Dual-Strategy Clustering Method. (a) Conceptual Representation of Methodology; (b) Overview of Data Preprocessing Steps.

Table 2
Dataset Overview-Left: Collective Anomaly over different practice centres; Right: Individual Anomaly for a selected Practice Centre over different processes.

Process	Practice Centre	Percentage (%)	Date	Process	Practice Centre	Percentage (%)	Date
DM3TT	E84075	31	01–07-2023	DP003	Y01066	1	08–08-2023
DM3TT	E85025	45	01–07-2023	LTA36	F84749	8	08–08-2023
DM3TT	E87046	39	01–07-2023	LT005	G85724	80	08–08-2023
DM3TT	E87063	34	01–07-2023	NDA19	Y00403	77	08–08-2023
DM3TT	F82034	42	01–07-2023	DMFOO	E85025	73	08–08-2023

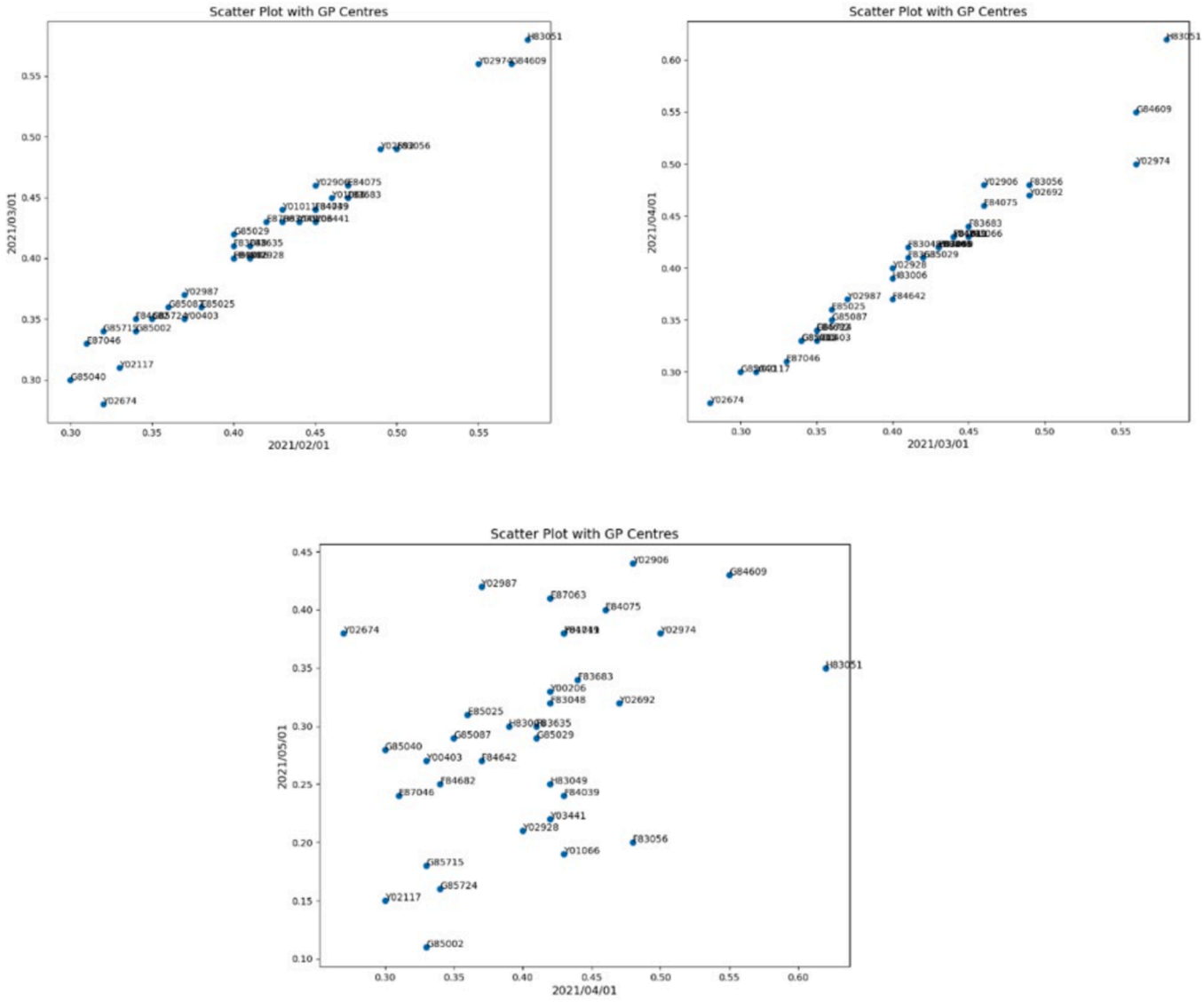


Fig. 2. Three samples of monthly patient percentage trends across GP centres showing normal and abnormal behaviours. Each diagram reports data for successive pairs of months over a 4-month period in 2021. In each plot, the horizontal axis represents the percentage of patients in a given month, while the vertical axis represents the percentage in the following month. Each point on the scatter plot represents a practice centre, with its position indicating the intersection of patient percentages across two successive months.

Table 3
Various combinations of tested hyperparameters.

Hyperparameter	Description	Tested Values	Optimal Value
epsilon (ε)	Maximum distance between two samples for one to be considered as in the neighbourhood of the other.	0.01, 0.02, 0.03, 0.04, 0.05	0.03
min_samples	Minimum number of points to form a dense region (cluster)	2, 3, 4, 5, 6	3

time series collectively exhibit anomalous behaviour and represent anomalies across multiple series.

This current work presents an approach based on time series analyses using unsupervised machine learning (ML). The dataset contains over 500 K of medical data records filtered for specific medical outcome processes from several tens of GP practice centres in the UK.

Typically, research on time series anomaly detection focuses on point/global anomalies within a series or subsequences as collective anomalies. In our work, data is examined from a unique perspective, considering entire time series as potential sources of anomalous patterns. This 'process-based' approach focuses on a single practice centre to study patterns of various processes within it. Identified abnormalities in these patterns can highlight specific time series as individual anomalies and prompt further investigation and corrective actions.

Our novel contribution to collective anomaly detection is a 'practice centre-based approach,' where each centre represents a time series showing the pattern of a process or group of processes over time (units of months). This approach enables centres to monitor changes compared to previous months, fostering better control and aiding the understanding of changes in a timely manner. It also allows for preventive actions if collective anomalies are detected.

Assessing the statistical significance of detected anomalies can be challenging within the context of a clustering-based approach because traditional accuracy metrics are not directly applicable. Furthermore, given the data is UK healthcare centric, non-UK healthcare data sets may require different pre-processing requirements. Our present work focusses on a novel anomaly detection method rather than expanding it

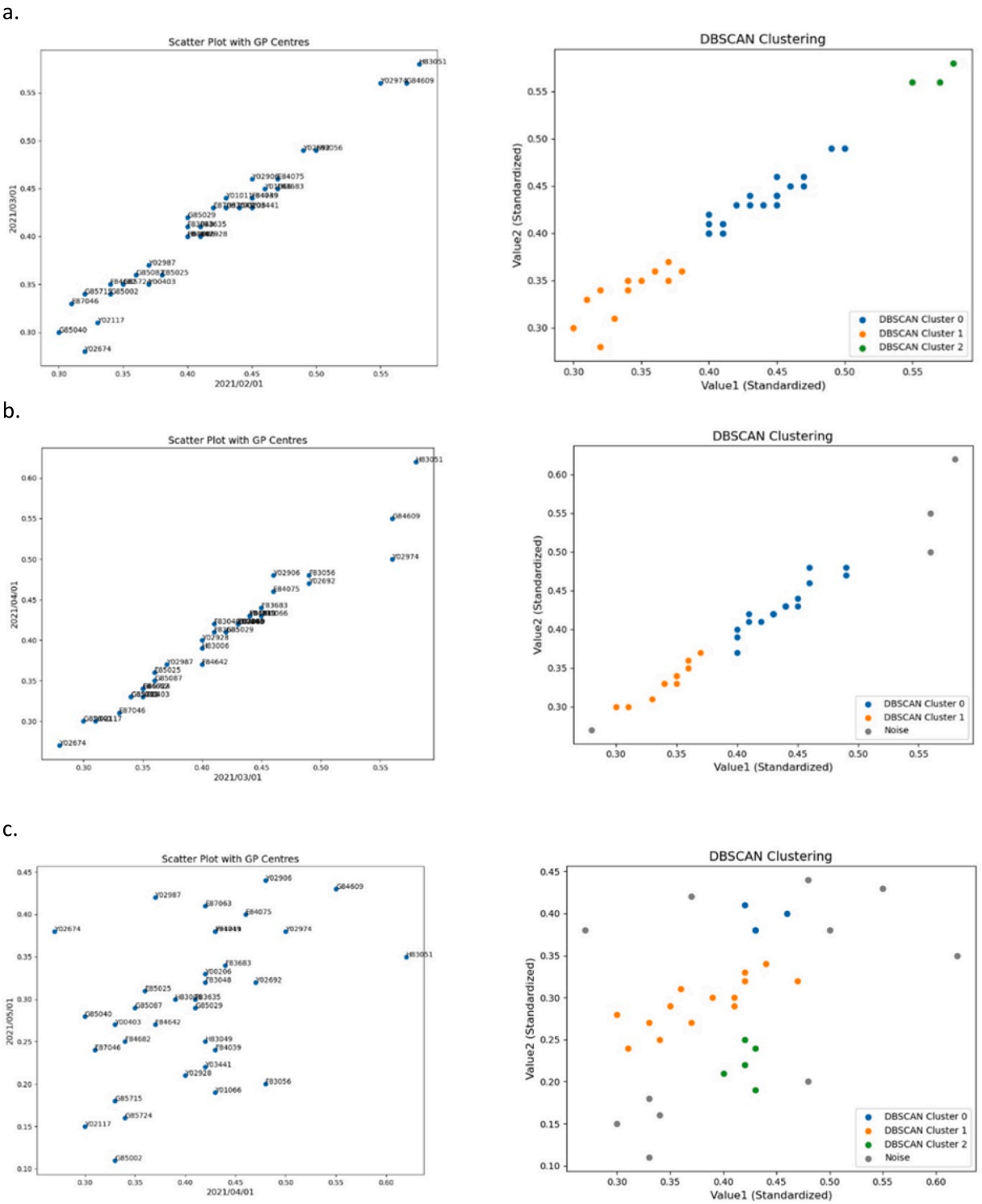
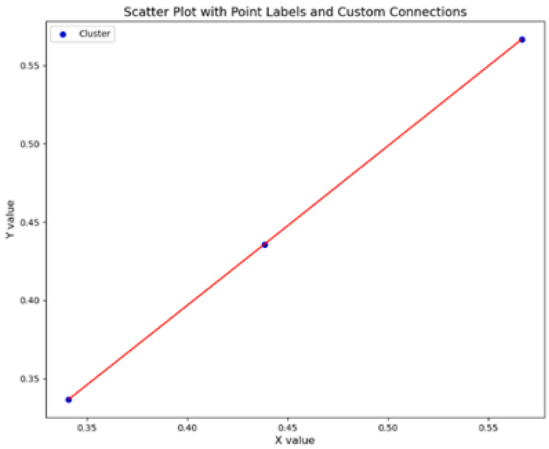


Fig. 3. Clustering results for the distribution of patients across GP centres between February and May 2021; 'a': February and March; 'b': March and April; 'c': April and May. In each case the left diagram presents the raw scatter data and the right diagram presents the cluster affiliations.

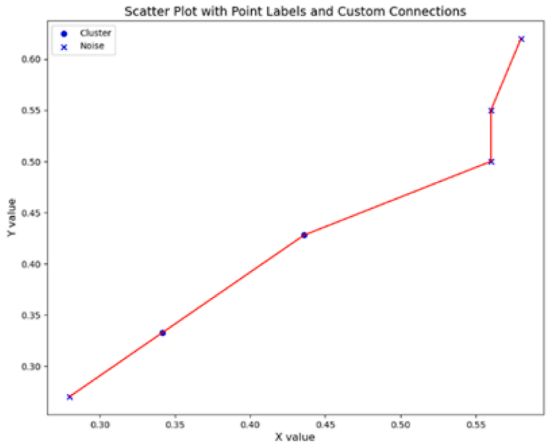
to include what would be detailed case/domain-specific interpretations and where applicable their remedial actions, of which for both there could be very many. We therefore recommend that any anomalies identified using our novel dual-strategy, clustering-specific method should be reviewed and confirmed by domain experts to assess their clinical relevance and to accurately label them as abnormal.

The most significant advantage of our unsupervised anomaly detection methods is the ability to identify anomalies that have not been previously identified by other means. Visual representations of the data are provided making it accessible and interpretable for non-technical, business-oriented users. The approach does not require any personal information thereby maintaining privacy even when non-GP personnel interact with the system. This overcomes the challenge posed by the absence of labelled data and offers a flexible and unsupervised solution to anomaly detection in healthcare.

a.



b.



c.

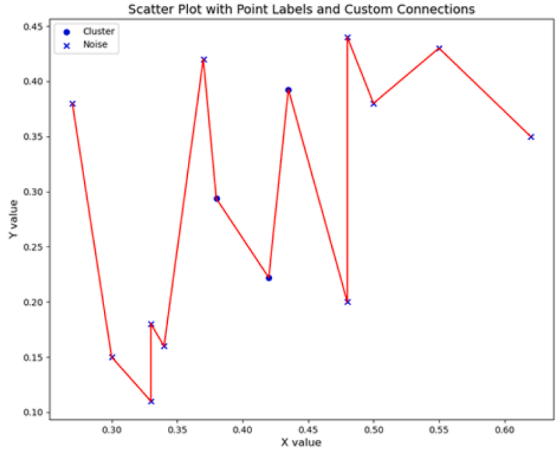


Fig. 4. Linear alignment of cluster centres and noises. 'a': shows the centres (means) of each cluster from Fig. 3.a, connected with straight lines. In 'b': the geometric centres of the two identified clusters from Fig. 3.b are calculated, and connections a re drawn to the four noise points. These centres serve as representative points that encapsulate the general characteristics of each cluster, and the connections highlight their relationship with the noise points, providing a comprehensive view of the data's structure. 'c': demonstrates the connections between the centres of the clusters and the outlier data points from Fig. 3.c, resulting in a complex and chaotic plot.

Table 4
Variability in time series lengths across different diseases.

Length	Outcome Process	Number
6	['LTA74′, 'LTA75′, 'LTA76′, 'LTA77′, 'PE0B5′, 'PE0B9′, 'PE0BA', 'PE0BB']	8
11	['DMS01′, 'LTA21′, 'LTA26′, 'LTA30′, 'LTA34′, 'LTA35′, 'LTA38′, 'LTA39′, 'LTA40′, 'LTA42′, 'LTA49′, 'LTA52′, 'MDS02′, 'NDA03′, 'PV038']	15
12	['LTA14′, 'LTA43′, 'NDA07']	3
13	['LTA46′, 'NDA04']	2
18	['DH001′, 'LTA64′, 'NDA01′, 'NDA10′, 'NDA12′, 'NDA14′, 'NDA15′, 'NDA16']	8
39	['LT012′, 'LT013′, 'LT014′, 'LT016′, 'LT018′, 'LTA53′, 'LTA68']	7
40	['COW32′, 'DM3TT', 'DM8CP', 'DM9CP', 'DMACR', 'DMBMI', 'DMBP1′, 'DMCHO', 'DMFOO', 'DMGFR', 'DMHBA', 'DMRET', 'DMSMO', 'LT001′, 'LT002′, 'LT003′, 'LT004′, 'LT005′, 'LT006′, 'LT007′, 'LT008′, 'LT010′, 'LT015′, 'LT017′, 'LT019′, 'LT023′, 'LT026′, 'LT027′, 'LT028′, 'LT029′, 'LT030′, 'LT040′, 'LT041′, 'LT042′, 'LT043′, 'LT044′, 'LT045′, 'LT046′, 'LT047′, 'LT050′, 'LT051′, 'LT052′, 'LT053′, 'LT054′, 'LT055′, 'LT056′, 'LT057′, 'LTA09′, 'LTA11′, 'LTA25′, 'LTA27′, 'NDA18′, 'NDA19′, 'NDA20′, 'X0Z4H']	55

Table 5
Descriptive overview of some medical processes.

Process Code	Description
DM3TT	The % of patients with Type 2 diabetes displaying in-year control for Hba1c, cholesterol and blood pressure in line with targets set out in the UK National Diabetes Audit programme.
DM8CP	The % of patients receiving a comprehensive diabetic assessment incorporating 8 key processes that all diabetic patients should receive annually (excluding diabetic retinal eye screening).
DM9CP	The % of patients receiving a comprehensive diabetic assessment incorporating 9 key processes that all diabetic patients should receive annually (including diabetic retinal eye screening).
DMACR	Urine test calculating albumin:creatinine ratio. This is an early marker of kidney disease.
LTA11	Referral to a structured education programme for newly diagnosed diabetic patients.
LTA27	Patients who have the wrong diabetic code in their medical record.

2. Materials and methods

Many methods have been used for anomaly detection (Table 1). For this research, DBSCAN and SOM were selected by the authors due to their appropriateness for dealing with medical data records[64,37]. These techniques form the core of a generalized conceptual procedure presented in this work for anomaly detection in healthcare data (Fig. 1 (a)).

2.1. Dataset

The dataset used comprises 549,511 records extracted from a medical database (AT Medics) for the period 2018–2023. All occurrences (a variable number per month per practice) for each 162 processes for every 37 practice centres have been extracted. Each occurrence value is the number of patients with that process at the practice centre of interest divided by the total number of registered patients across all centres, expressed as a percentage (Table 2).

2.2. Data preprocessing

The focus of the data in this study is on processes conducted within GP centres, specifically examining the status of these processes without considering patient demographics or personal details. This eliminates the need for handling sensitive or personal information thus avoiding privacy concerns. Since we do not deal with multiple features standard data cleaning procedures such as encoding categorical variables or imputing missing values do not apply. In fact, the dataset contains a single time series variable that indicates whether all patients with a particular condition (e.g., type 2 diabetes) regardless of demographic details have undergone a specific process. Since different processes may occur at different times across various centres there are variations in the number of records for different practice centres within a month, ranging from multiple records to a single record. An appropriate preprocessing chain is implemented to ensure a consistent time series dataset (Fig. 1 (b)).

The important preprocessing steps are summarized as follows:

- The mean value for each month's records is computed and recorded on the first day to ensure consistency and comparability across all practice centres.
- To identify a collective anomaly, a specific process (DM3TT) is chosen, and the dataset grouped by different practice centres. This dataset comprises a subset of 3,500 observations, representing the percentage of patients with Type 2 diabetes who achieved in-year control of HbA1c, cholesterol, and blood pressure in accordance with targets set by the UK National Diabetes Audit programme.
- To identify an individual anomaly, an inverse approach is taken, where data from a single anonymous GP centre is grouped across all processes. Each process within that particular GP centre is treated as an individual time series, resulting in 162 separate time series spanning the entire duration of the data collection.

2.3. Methods

2.3.1. DBSCAN clustering

Density-based clustering algorithms like DBSCAN (Ester et al., 1996) effectively identify areas with high data concentration and isolating outliers. Monthly clustering enables a comparison of consecutive months and provides a valuable temporal dimension to anomaly detection. It is flexible in that it can group similar objects with arbitrary shapes while isolating outliers that lack close neighbours.

The intuitive concept behind DBSCAN is that a data point is considered part of a cluster if it has a sufficient number of close neighbours surrounding it. The different points in DBSCAN are classified based on the concept of the epsilon neighbourhood of a point using equation (1).

$$N_\varepsilon(p) = \{q | d(p,q) < \varepsilon\} \tag{1}$$

where:

$N_\varepsilon(p)$ is the epsilon neighbourhood of point $.p$
$d(p,q)$ represents the Euclidean distance between two points.
ε is the radius.

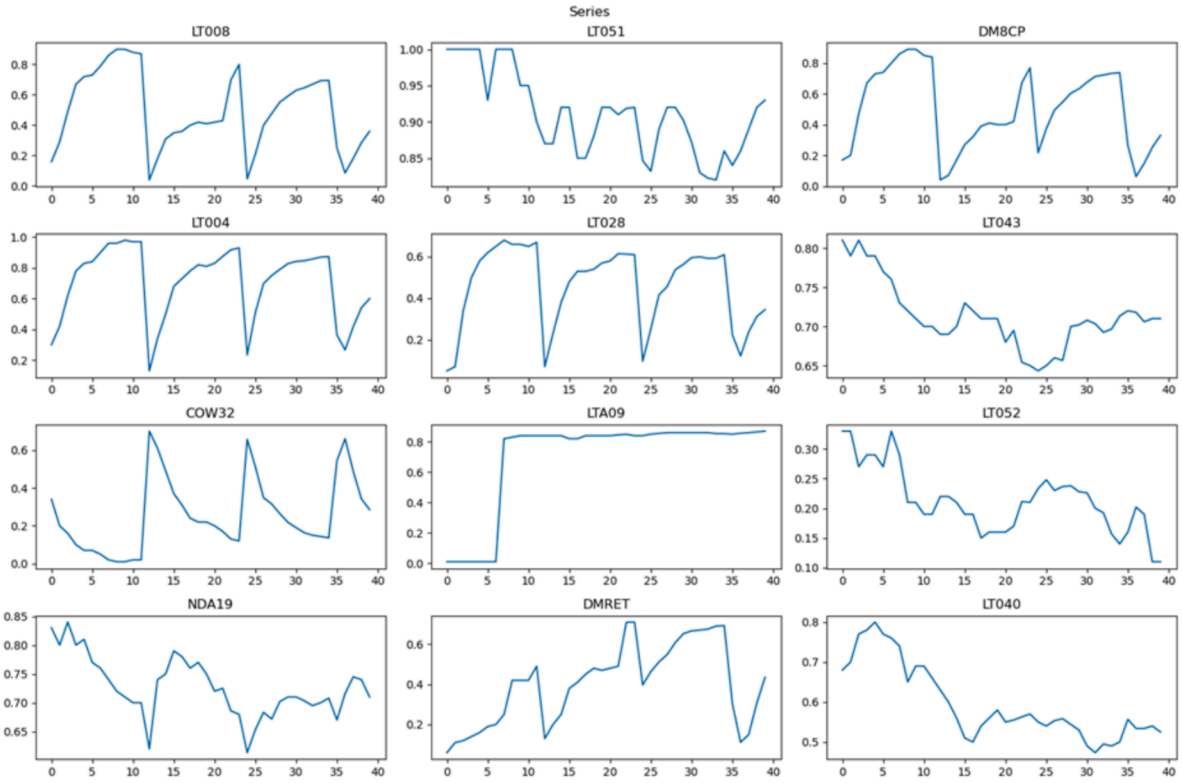
If a set contains at least the predefined minimum number of points, it is referred to as a core point. If the set contains fewer than the minimum but more than one point, it is considered a border point. Conversely, if the set does not contain any other points, it is classified as an outlier. The algorithm operates based on the symmetric principle of direct reachability. Specifically, a point p is directly density-reachable from a point q with respect to ϵ and the minimum number of points if p is within the ϵ-neighbourhood of q (denoted as $N_\varepsilon(q)$) and the number of points in $N_\varepsilon(q)$, denoted as $| N_\varepsilon(q)|$, is greater than or equal to the minimum number of points.

The DBSCAN algorithm starts by selecting a core point to form the first cluster, adding all nearby core points. It then includes nearby border points, which do not further expand the cluster. Other clusters are formed similarly, with remaining points classified as outliers.

2.3.2. SOM clustering

SOM (Self-Organizing Map) is an unsupervised learning algorithm

a.



b.

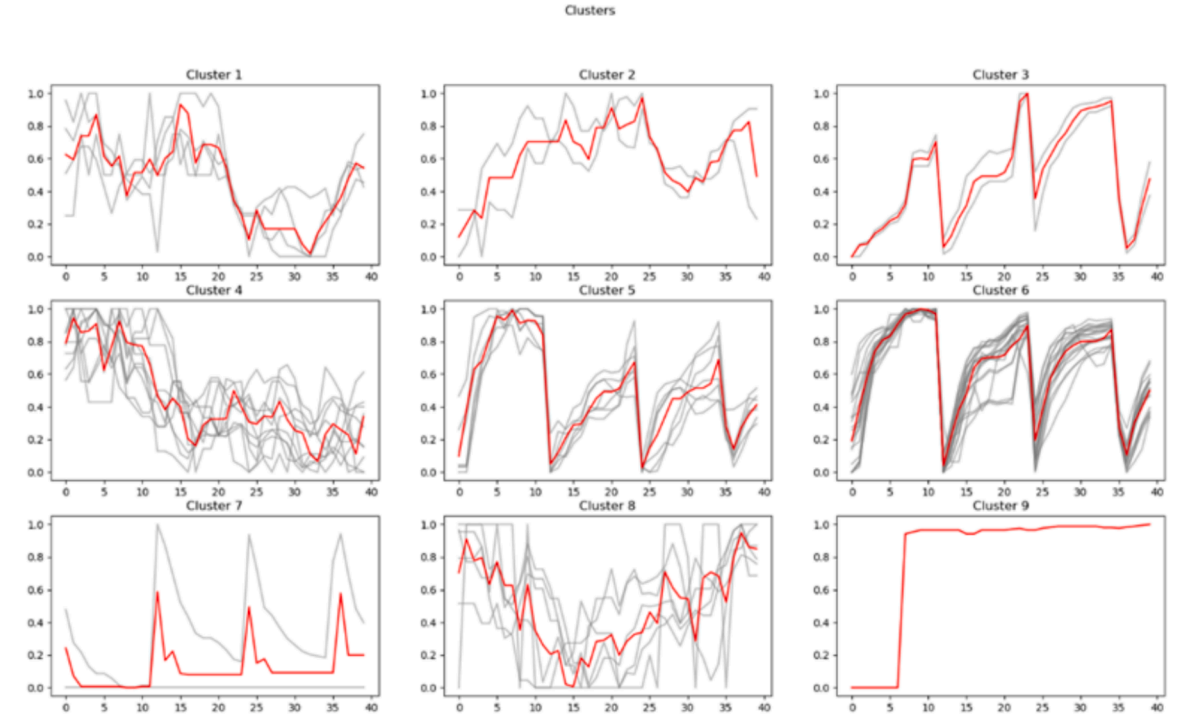


Fig. 5. Graphics of Time Series data. 'a': Time series plots of 12 processes for a GP centre since 2018; 'b': Clusters of time series processes for a GP Centre Since 2018; 'c': Distribution of the time series in clusters.

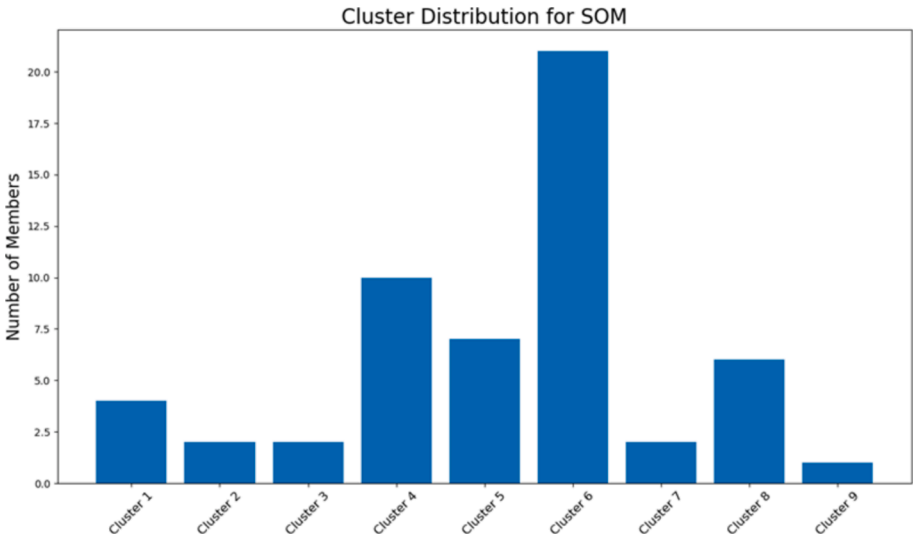


Fig. 5. (*continued*).

for clustering, mapping similar data points to nearby grid locations and dissimilar points to distant ones[24]. It operates in three stages: competitive, cooperative, and adaptation[34]. In the competitive stage, the Best Matching Unit (BMU) is identified based on similarity. During the cooperative stage, the BMU influences its neighbouring neurons through a neighbourhood function (equation (2):

$$h_{ij}(x,t) = \exp(-\frac{{d_{ij}}^2}{2\sigma^2(t)}) \tag{2}$$

Here, d_{ij} is the distance between neurons i and j, and $\sigma(t)$ represents the neighbourhood radius at time t. As expected, the influence of the neighbourhood function diminishes for neurons that are farther apart and tends to zero, which is evident from the function's form. Also, as the number of iterations increases over time, the influence of the neighbourhood function is expected to decrease because $\sigma(t)$ decreases with time according to equation (3):

$$\sigma(t) = \sigma_0 \exp(-\frac{t}{\tau}) \tag{3}$$

Where σ_0 represents the initial neighbourhood radius, t signifies the current time, and τ is a time constant parameter that determines the rate at which the neighbourhood radius decreases over time.

This indicates that as the learning process proceeds and more iterations occur, the neighbourhood radius shrinks. The final phase is adaptation, where the winning neuron and its neighbours' weights are updated to move closer to the input data. The weight update is governed by equation (4):

$$\omega_i(t+1) = \omega_i(t) + \eta h_{ij}(x,t)(x - \omega_i) \tag{4}$$

Where η represents the learning rate, which also decreases over time. This is to ensure that if the data is not an outlier, it will also be represented in the earlier stages of training.

When clustering time series data using SOM, traditional similarity measures like Euclidean distance may be inappropriate due to potential time shifts and varying speeds within the patterns. Dynamic Time Warping (DTW) is a more suitable measure, as it accounts for temporal distortions, shifts, and speed variations[44,48].

3. Results and discussion

3.1. Collective anomaly

Scatter plots for all GP practice centres for every two consecutive months between 2018 and 2023 were generated. Anomalies do not necessarily occur every month and the examples selected were chosen as these months exhibit notable unusual patterns (Fig. 2).

The GP centres typically exhibit normal behaviour when they are positioned on or near the diagonal bisector. This pattern is expected, as significant differences between two successive months are uncommon. Consequently, GP centres can be categorized into distinct clusters along the bisector line. Centres with a larger size (more patients) tend to be situated in the upper right quadrant of the plot.

To determine the optimal number of clusters for each month's clustering analysis, the grid search method was employed to identify the best epsilon and minimum points for DBSCAN clustering (Table 3).

The selected hyperparameters were those that consistently produced the most meaningful and well-separated clusters across multiple iterations based on the Silhouette score.

In Fig. 3.a, the practice centres form three distinct clusters are aligned along the same straight line. Repeating this visualization for comparable data for March and April 2021 using DBSCAN reveals two distinct clusters and four outliers (noises), as illustrated in Fig. 3.b. Finally, Fig. 3.c presents the clustering results for April and May 2021, which were selected for their most obvious chaotic and anomalous behaviour. In this case, the grouping forms four distinct clusters accompanied by eight outliers, showing a striking deviation of data points from the bisector line.

Fig. 4 provides further insights by visualizing the representative points and their connections. Upon comparing Fig. 4.b with Fig. 4.a, clear differences in the clustering pattern are evident. Specifically, two GP centres located in the top right and one GP centre from the middle cluster have undergone a significant shift. These centres, previously part of their respective clusters, have now diverged and are behaving as three outliers in the top right of Fig. 4.b. Additionally, one GP centre from the lower cluster in Fig. 4.a has also separated, now appearing as a single outlier. This shift in positioning indicates both point anomalies and collective anomalies, signalling a need for further investigation into the behaviours of these specific practice centres. The high level of variability and irregularity in Fig. 4.c could be indicative of underlying issues or anomalies within the dataset. Overall, the visualizations show that over time clusters tend to break apart into outliers or form new clusters. This shared behaviour across multiple GP centres qualifies as a collective

anomaly.

Studies by Cerqueira et. al. [7], Li et. al. (2021), Keogh et. al. [20] and Röhrig [42] emphasize the importance of subsequence partitioning and try to identify specific types of subsequences, such as time series discords or anomalous segments within medical data. Additionally, Li et. al. [29] proposes a unified model for detecting both collective (as subsequences) and point anomalies. Instead of solely considering individual subsequences, the approach proposed in this current work analyses the collective behaviour of multiple data points across different time series, offering a more comprehensive perspective and effectively surpassing the limitations associated with focusing solely on subsequences.

3.2. Individual anomaly

In this section only the data from one specific GP centre for all 162 distinct outcome processes between 2018–2023 are considered. The initial step reviews the lengths of some individual time series within the dataset, revealing inconsistencies (Table 4).

A description of some codes of Table 4 as outcomes in medical processes are presented in Table 5.

To maintain comparability and robustness in the subsequent analyses and to ensure that all selected time series were equipped with the same comprehensive temporal context, the authors selected the longest 55 time series each characterized by a consistent length of 40 data points, the maximum possible in our dataset, for graphing and inspection.

Fig. 5 (a, b and c) offer three visual overviews that aid in detecting potential outliers or anomalies. A subset of 12 members from selected time series were plotted in Fig. 5.a to show that while certain time series can be grouped and clustered together due to their similar patterns, some irregular patterns, such as LTA09, appear to be anomalous in deviating from this trend. Fig. 5.b presents the outcome of the SOM analysis and reveals nine distinct clusters, with the red series representing the average time series within each cluster. Notably, some clusters could be anomalous in comprising only a single time series or a small number (two or three). Fig. 5.c provides a breakdown of the number of time series within each cluster. The most populated cluster is Cluster 6, containing 21 time series. In contrast, Cluster 9 is unique, comprising only a single time series and clusters 2, 3, and 7 each contain just two time series. These clusters with fewer time series may be anomalous. These three visualizations are of special interest as they may contain anomalous time series that deviate from commonly detected patterns and thus require closer examination.

Our focus is on process-based anomaly detection when it comes to detecting individual anomalies. Individual medical outcome processes presented as separate time series within a single practice centre have been examined. This approach enables us to group processes into different clusters, where a cluster with one or two outlier members prompts further inquiry into the causes and potential solutions to address these irregularities.

4. Conclusions

The exploration of anomaly detection in the UK's NHS holds substantial implications for the enhancement of healthcare quality and the pre-emptive identification of potential crises both nationally and further afield.

In the practice centre-based analysis, the behaviour of various centres can be examined by investigating disease trends over time. When data points (representing practice centres) align along a straight line on a successive month bivariate scattergram it typically indicates expected behaviour, as large variations between the two months are relatively rare. However, deviations from this pattern – particularly the emergence of new clusters or outliers – necessitate further investigation (collective anomalies). In the process-based analysis, individual diseases presented as separate time series within a single practice centre are analysed. The detection of anomalies within these series, such as an unexpected pattern in disease incidence, also prompts further inquiry (individual anomalies). These anomalies could signify underlying shifts in patient dynamics or healthcare delivery that require attention and potential solutions to address these irregularities.

Unlike traditional statistical methods, which either struggle with high-dimensional data or require extensive training, our approach excels in identifying both collective and individual anomalies. Traditional methods often focus on point or subsequence anomalies within a single time series, while our method detects anomalies across multiple time series, offering a more comprehensive perspective. By leveraging DBSCAN's ability to identify clusters and outliers and SOM's capacity to preserve data relationships, our method addresses limitations of existing techniques such as handling varying cluster shapes and avoiding the need for pre-labelled data. Additionally, our approach provides accessible visual representations, enhancing interpretability for non-technical users.

Our research suggests that the dual-strategy approach for monitoring healthcare time series data by unsupervised learning is a useful structured approach for anomaly detection. Whilst this is effective in identifying irregular patterns, it is essential to interpret these findings within the broader clinical context. What may be statistically anomalous may not necessarily be clinically relevant, and vice versa. Therefore, collaboration with healthcare professionals is crucial to align detection with clinical significance.

5. Summary table

What was already known on the topic.

- Healthcare anomaly detection is critical for identifying operational inefficiencies, fraud and emerging health issues, with previous research emphasising the importance of early warnings and timely interventions.
- Time series analysis in healthcare anomaly detection has often focused on single-point or subsequence anomalies within individual time series, missing broader, collective anomalies across multiple related series.
- Privacy and data sensitivity are significant challenges in healthcare anomaly detection, as traditional models often require large, labelled datasets or personal data, which are not always accessible or desirable due to confidentiality concerns.

What this study added to our knowledge.

- Focusing on entire time series as sources of anomalies rather than only point anomalies, facilitating a more holistic approach to healthcare anomaly detection.
- The study achieves anomaly detection that preserves data privacy, making it suitable for sensitive healthcare data without requiring personal information.
- Visual interpretability for non-technical users, allowing healthcare practitioners to identify anomalies quickly and enabling proactive decision-making without needing technical expertise.

CRediT authorship contribution statement

Farbod Khanizadeh: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Alireza Ettefaghian:** Writing – review & editing, Validation, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **George Wilson:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Amirali Shirazibeheshti:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Tarek Radwan:**

Validation, Resources, Funding acquisition, Conceptualization. **Cristina Luca:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Amirali Shirazibeheshti, Tarek Radwan reports financial support was provided by AT Medics Ltd. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2024.105696.

References

[1] R. Bauder, T.M. Khoshgoftaar, N. Seliya, A survey on the state of healthcare upcoding fraud analysis and detection, *Health Serv. Outcomes Res. Method.* 17 (2017) 31–55, https://doi.org/10.1007/s10742-016-0154-8.

[2] S.E. Benkabou, K. Benabdeslem, B. Canitia, Unsupervised outlier detection for time series by entropy and dynamic time warping, *Knowl. Inf. Syst.* 54 (2) (2018) 463–486, https://doi.org/10.1007/s10115-017-1067-8.

[3] Bergman, L. and Hoshen, Y., 2020. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*. Doi: 10.48550/arXiv.2005.02359.

[4] R.G. Brereton, Self organising maps for visualising and modelling, *Chem. Cent. J.* 6 (2) (2012) 1–15.

[5] P.L. Brockett, L.L. Golden, J. Jang, C. Yang, A comparison of neural network, statistical methods, and variable choice for life insurers' financial distress prediction, *Journal of Risk and Insurance* 73 (3) (2006) 397–419.

[6] M. Çelik, F. Dadaşer-Çelik, A.Ş. Dokuz, Anomaly detection in temperature data using DBSCAN algorithm, in: 2011 International Symposium on Innovations in Intelligent Systems and Application, IEEE, 2011, pp. 91–95.

[7] V. Cerqueira, L. Torgo, C. Soares, Early anomaly detection in time series: a hierarchical approach for predicting critical health episodes, *Mach. Learn.* (2023) 1–22, https://doi.org/10.1007/s10994-022-06300-x.

[8] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, *ACM Computing Surveys (CSUR)* 41 (3) (2009) 1–58.

[9] L. Duan, L. Xu, Y. Liu, J. Lee, Cluster-based outlier detection, *Ann. Oper. Res.* 168 (2009) 151–168.

[10] T. Ekina, F. Leva, F. Ruggeri, R. Soyer, Application of bayesian methods in detection of healthcare fraud, *Chemical Engineering Transaction* 33 (2013).

[11] G. Erdogan, Spectral methods for outlier detection in machine learning, Bougazixci University, 2012. Master report.

[12] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, C. Fookes, Deep learning for medical anomaly detection–a survey, *ACM Computing Surveys (CSUR)* 54 (7) (2021) 1–37, https://doi.org/10.1145/3464423.

[13] R. Foorthuis, On the nature and types of anomalies: a review of deviations in data, *International Journal of Data Science and Analytics* 12 (4) (2021) 297–331, https://doi.org/10.1007/s41060-021-00265-1.

[14] M. Goldstein, S. Uchida, A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, *PLoS One* 11 (4) (2016) e0152173.

[15] R. Griffith, Fraud in the NHS, *Br. J. Nurs.* 28 (19) (2019) 1268–1269.

[16] F.E. Grubbs, Procedures for detecting outlying observations in samples, *Technometrics* 11 (1) (1969) 1–21.

[17] R. Ikono, O. Iroju, J. Olaleke, T. Oyegoke, Meta-analysis of fraud, waste and abuse detection methods in healthcare, *Niger. J. Technol.* 38 (2) (2019) 490–502.

[18] P.K. Jain, M.S. Bajpai, R. Pamula, A modified DBSCAN algorithm for anomaly detection in time-series data with seasonality, *Int. Arab J. Inf. Technol.* 19 (1) (2022) 23–28.

[19] V.O. Kayhan, M. Agrawal, S. Shivendu, Cyber threat detection: Unsupervised hunting of anomalous commands (UHAC), *Decis. Support Syst.* 168 (2023) 113928, https://doi.org/10.1016/j.dss.2023.113928.

[20] E. Keogh, J. Lin, S.H. Lee, H.V. Herle, Finding the most unusual time series subsequence: algorithms and applications, *Knowl. Inf. Syst.* 11 (2007) 1–27.

[21] E. Keogh, S. Lonardi, B.Y.C. Chiu, Finding surprising patterns in a time series database in linear time and space, in: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 550–556.

[22] Khan, K., Rehman, S.U., Aziz, K., Fong, S. and Sarasvady, S., 2014, February. DBSCAN: Past, present and future. In *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)* (pp. 232-238). IEEE.

[23] M. Kirlidog, C. Asuk, A fraud detection approach with data mining in health insurance, *Procedia Soc. Behav. Sci.* 62 (2012) 989–994.

[24] T. Kohonen, The self-organizing map, *Proc. IEEE* 78 (9) (1990) 1464–1480.

[25] T. Kohonen, Essentials of the self-organizing map, *Neural Netw.* 37 (2013) 52–65.

[26] T.T. Kuo, A. Pham, Detecting model misconducts in decentralized healthcare federated learning, Int. J. Med. Inf. 158 (2022) 104658, https://doi.org/10.1016/j.ijmedinf.2021.104658.

[27] W. Lee, D. Xiang, Information-theoretic measures for anomaly detection, in: Proceedings 2001 IEEE Symposium on Security and Privacy. s&p, IEEE, 2000, pp. 130–143.

[28] J. Li, H. Izakian, W. Pedrycz, I. Jamal, Clustering-based anomaly detection in multivariate time series data, *Appl. Soft Comput.* 100 (2021) 106919, https://doi.org/10.1016/j.asoc.2020.106919.

[29] Z. Li, Z. Xiang, W. Gong, H. Wang, Unified model for collective and point anomaly detection using stacked temporal convolution networks, *Appl. Intell.* 52 (3) (2022) 3118–3131, https://doi.org/10.1007/s10489-021-02559-0.

[30] Liu, Q. and Vasarhelyi, M., 2013, November. Healthcare fraud detection: A survey and a clustering model incorporating geo-location information. In *29th world continuous auditing and reporting symposium (29WCARS), Brisbane, Australia.*

[31] Luo, W. and Gallagher, M., 2010, December. Unsupervised DRG upcoding detection in healthcare databases. In *2010 IEEE International Conference on Data Mining Workshops* (pp. 600-605). IEEE.

[32] M.C. Massi, F. Ieva, E. Lettieri, Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases, *BMC Med. Inf. Decis. Making* 20 (2020) 1–11.

[33] A. Mehbodniya, I. Alam, S. Pande, R. Neware, K.P. Rane, M. Shabaz, M. V. Madhavan, Financial fraud detection in healthcare using machine learning and deep learning techniques, *Secur. Commun. Netw.* 2021 (2021) 1–8, https://doi.org/10.1155/2021/9293877.

[34] D. Miljković, Brief review of self-organizing maps, in: 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2017, pp. 1061–1066.

[35] S. Monalisa, F. Kurnia, Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour, *Telkomnika (telecommunication Computing Electronics and Control)* 17 (1) (2019) 110–117.

[36] B. Murdoch, Privacy and artificial intelligence: challenges for protecting health information in a new era, *BMC Med. Ethics* 22 (1) (2021) 1–5, https://doi.org/10.1186/s12910-021-00687-3.

[37] Y.A. Nanehkaran, Z. Licai, J. Chen, A.A. Jamel, Z. Shengnan, Y.D. Navaei, M. A. Aghbolagh, Anomaly detection in heart disease using a density-based unsupervised approach, *Wirel. Commun. Mob. Comput.* 2022 (2022), https://doi.org/10.1155/2022/6913043.

[38] E.W. Ngai, Y. Hu, Y.H. Wong, Y. Chen, X. Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, *Decis. Support Syst.* 50 (3) (2011) 559–569.

[39] A. Nowak-Brzezińska, C. Horyń, Self-Organizing Map algorithm as a tool for outlier detection, *Procedia Comput. Sci.* 207 (2022) 2162–2171, https://doi.org/10.1016/j.procs.2022.09.276.

[40] R. Pamula, J.K. Deka, S. Nandi, An outlier detection method based on clustering, in: 2011 Second International Conference on Emerging Applications of Information Technology, IEEE, 2011, pp. 253–256.

[41] T. Pourhabibi, K.L. Ong, B.H. Kam, Y.L. Boo, Fraud detection: A systematic literature review of graph-based anomaly detection approaches, *Decis. Support Syst.* 133 (2020) 113303, https://doi.org/10.1016/j.dss.2020.113303.

[42] Röhrig, R., 2021, June. Semantic anomaly detection in medical time series. In *German Medical Data Sciences: Bringing Data to Life: Proceedings of the Joint Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology (gmds EV) and the Central European Network-International Biometric Society (CEN-IBS) 2020 in Berlin, Germany* (Vol. 278, p. 118). IOS Press.

[43] E. Šabić, D. Keeley, B. Henderson, S. Nannemann, Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data, *AI & Soc.* 36 (1) (2021) 149–158, https://doi.org/10.1007/s00146-020-00985-1.

[44] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans. Acoust. Speech Signal Process.* 26 (1) (1978) 43–49.

[45] O. Salem, Y. Liu, A. Mehaoua, R. Boutaba, Online anomaly detection in wireless body area networks for reliable healthcare monitoring, *IEEE J. Biomed. Health Inform.* 18 (5) (2014) 1541–1551.

[46] J.K. Samriya, S. Kumar, S. Singh, Efficient K-means clustering for healthcare data, *Advanced Journal of Computer Science and Engineering (AJCST)* 4 (2) (2016) 1–7.

[47] D. Samariya, J. Ma, S. Aryal, X. Zhao, Detection and explanation of anomalies in healthcare data, *Health Inf. Sci. Syst.* 11 (1) (2023) 20, https://doi.org/10.1007/s13755-023-00221-2.

[48] P. Senin, Dynamic time warping algorithm review, *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* 855 (1–23) (2008) 40.

[49] K. Sheridan, T.G. Puranik, E. Mangortey, O.J. Pinon-Fischer, M. Kirby, D.N. Mavris, An application of dbscan clustering for flight anomaly detection during the approach phase, in: *AIAA Scitech 2020 Forum*, 2020, p. 1851.

[50] H. Shin, H. Park, J. Lee, W.C. Jhee, A scoring model to detect abusive billing patterns in health insurance claims, *Expert Syst. Appl.* 39 (8) (2012) 7441–7450.

[51] A. Shirazibeheshti, A. Ettefaghian, F. Khanizadeh, G. Wilson, T. Radwan, C. Luca, Automated detection of patients at high risk of polypharmacy including anticholinergic and sedative medications, *Int. J. Environ. Res. Public Health* 20 (12) (2023) 6178.

[52] M.L. Shyu, S.C. Chen, K. Sarinnapakorn, L. Chang, A novel anomaly detection scheme based on principal component classifier, in: Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, IEEE Press, 2003, pp. 172–179.

[53] T. Sipes, S. Jiang, K. Moore, N. Li, H. Karimabadi, J.R. Barr, Anomaly detection in healthcare: Detecting erroneous treatment plans in time series radiotherapy data, *International Journal of Semantic Computing* 8 (03) (2014) 257–278.
[54] Sowah, R.A., Kuuboore, M., Ofoli, A., Kwofie, S., Asiedu, L., Koumadi, K.M. and Apeadu, K.O., 2019. Decision support system (DSS) for fraud detection in health insurance claims using genetic support vector machines (GSVMs). *Journal of Engineering*, *2019*.
[55] I. Steinwart, D. Hush, C. Scovel, A classification framework for anomaly detection, *J. Mach. Learn. Res.* 6 (2) (2005).
[56] N. Stylianou, R. Fackrell, C. Vasilakis, Are medical outliers associated with worse patient outcomes? A retrospective study within a regional NHS hospital using routine data, *BMJ Open* 7 (5) (2017) e015676.
[57] Sun, P., Chawla, S. and Arunasalam, B., 2006, April. Mining for outliers in sequential databases. In *Proceedings of the 2006 SIAM international conference on data mining* (pp. 94-105). Society for Industrial and Applied Mathematics.
[58] Suresh, N.C., De Traversay, J., Gollamudi, H., Pathria, A.K. and Tyler, M.K., Fair Isaac Corp, 2014. *Detection of upcoding and code gaming fraud and abuse in prospective payment healthcare systems*. U.S. Patent 8,666,757.
[59] M. Tang, B.S.U. Mendis, D.W. Murray, Y. Hu, A. Sutinen, Unsupervised fraud detection in Medicare Australia, in: In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, 2011, pp. 103–110.
[60] D. Thornton, G. van Capelleveen, M. Poel, J. van Hillegersberg, R.M. Mueller, April. Outlier-based Health Insurance Fraud Detection for US Medicaid Data, In *ICEIS* 2 (2014) 684–694.
[61] S. Thudumu, P. Branch, J. Jin, J. Singh, A comprehensive survey of anomaly detection techniques for high dimensional big data, *Journal of Big Data* 7 (2020) 1–30, https://doi.org/10.1186/s40537-020-00320-x.
[62] R.S. Tsay, D. Pena, A.E. Pankratz, Outliers in multivariate time series, *Biometrika* 87 (4) (2000) 789–804.
[63] G. van Capelleveen, M. Poel, R.M. Mueller, D. Thornton, J. van Hillegersberg, Outlier detection in healthcare fraud: A case study in the Medicaid dental domain, *Int. J. Account. Inf. Syst.* 21 (2016) 18–31.
[64] H. Wang, S. He, T. Liu, Y. Pang, J. Lin, Q. Liu, K. Han, J. Wang, G. Jeon, QRS detection of ECG signal using U-Net and DBSCAN, *Multimed. Tools Appl.* 81 (10) (2022) 13319–13333, https://doi.org/10.1007/s11042-021-10994-x.
[65] Q. Wang, B. Yan, H. Su, H. Zheng, in: March. Anomaly Detection for Time Series Data Stream, IEEE, 2021, pp. 118–122.
[66] Wei, Q., Ren, Y., Hou, R., Shi, B., Lo, J.Y. and Carin, L., 2018, February. Anomaly detection for medical images based on a one-class classification. In *Medical Imaging 2018: Computer-Aided Diagnosis* (Vol. 10575, pp. 375-380). SPIE.
[67] S. Wu, S. Wang, Information-theoretic outlier detection for large-scale categorical data, *IEEE Trans. Knowl. Data Eng.* 25 (3) (2011) 589–602.
[68] C. Zhang, X. Xiao, C. Wu, Medical fraud and abuse detection system based on machine learning, *Int. J. Environ. Res. Public Health* 17 (19) (2020) 7265, https://doi.org/10.3390/ijerph17197265.