REGULAR ARTICLE   **Open Access**

Forecasting patient flows with pandemic induced concept drift using explainable machine learning

Teo Susnjak[1]* and Paula Maddigan[1]

*Correspondence:
t.susnjak@massey.ac.nz
[1]School of Mathematical and
Computational Sciences, Massey
University, Auckland, New Zealand

Abstract

Accurately forecasting patient arrivals at Urgent Care Clinics (UCCs) and Emergency Departments (EDs) is important for effective resourcing and patient care. However, correctly estimating patient flows is not straightforward since it depends on many drivers. The predictability of patient arrivals has recently been further complicated by the COVID-19 pandemic conditions and the resulting lockdowns.

This study investigates how a suite of novel quasi-real-time variables like Google search terms, pedestrian traffic, the prevailing incidence levels of influenza, as well as the COVID-19 Alert Level indicators can both generally improve the forecasting models of patient flows and effectively adapt the models to the unfolding disruptions of pandemic conditions. This research also uniquely contributes to the body of work in this domain by employing tools from the eXplainable AI field to investigate more deeply the internal mechanics of the models than has previously been done.

The Voting ensemble-based method combining machine learning and statistical techniques was the most reliable in our experiments. Our study showed that the prevailing COVID-19 Alert Level feature together with Google search terms and pedestrian traffic were effective at producing generalisable forecasts. The implications of this study are that proxy variables can effectively augment standard autoregressive features to ensure accurate forecasting of patient flows. The experiments showed that the proposed features are potentially effective model inputs for preserving forecast accuracies in the event of future pandemic outbreaks.

Keywords: Forecasting; Patient flows; Machine learning; Explainable AI; Interpretable machine learning; Concept drift

1 Introduction

Urgent Care Clinics (UCCs) and Emergency Departments (EDs) are frontline healthcare providers, provisioning continuous care for a range of acute medical presentations. For a significant proportion of patients, UCCs and EDs are the initial pathways en route towards subsequent healthcare services and are thus prone to significant periodic congestion. The effects of delayed treatments at EDs due to overcrowding have been well investigated and research shows that these delays have been associated with negative clinical outcomes [42]. Although UCCs typically cater to less acute medical conditions, studies have also

noted that insufficient staff availability at these facilities could lead to serious conditions being overlooked [2] in the presence of overcrowding.

Therefore it is important to implement strategies which assist with planning and the effective allocation of human resources to manage patient flows. Forecasting patient arrivals is one such strategy which can enable a greater optimisation of medical-service resourcing at facilities situated beyond primary care providers. Recently, there has been an increased academic interest in developing forecasting models for solving this problem. Forecasting of this kind can operate at different levels of granularity, such as estimating total daily patient volumes or even higher frequencies, like hourly patient arrivals. The purpose of this study is to develop daily patient-flow forecasting models using machine learning for two large UCCs providing services to patients experiencing sudden illness or accident-related injury.

Forecasting patient arrival volumes to a level of accuracy which makes the models useful is non-trivial due to numerous influencing factors as well as underlying stochastic processes. Producing reliable patient flow estimates has also been further complicated recently with the outbreak of the COVID-19 pandemic which has contributed an additional level of complexity due to the disruption to usual patient flow patterns [25, 53]. These dislocations from otherwise stationary patterns in the dependent variable are termed *concept drift*. This phenomenon is encountered when the variable being forecasted no longer corresponds to the original data used for creating the model, ultimately resulting in the deterioration of accuracies. While these drifts can be sudden or gradual, ultimately a successful adaptive strategy depends on the ability to formulate an approach that captures the generative factors of the drifts as much as possible [43]. To date, no studies have yet emerged providing an analysis of how the concept drift arising from the COVID-19 pandemic and the resulting lockdown mandates have affected the patient-flow forecasting models, nor how the mitigation strategies can help adapt the forecasting models to increased volatility.

The current body of literature has primarily focused on forecasting solutions for EDs, with UCCs being largely overlooked; though the same drivers of patient demand and challenges apply to both UCCs and EDs. Most of the studies have concentrated on estimating total daily patient arrivals for the following day [5, 21, 32, 39, 49, 51, 53]. A few studies have investigated the feasibility of generating accurate forecasts as far as 30 days ahead [7, 29] and some 90 days ahead [28], which entails an even greater level of uncertainty and challenge due to compounding errors. For the most part, earlier studies relied on standard statistical methods for this type of modelling [2, 5, 9, 26]. Over time, machine learning approaches emerged [51, 53], including deep learning techniques [21, 39, 42]. As the sophistication of the underlying algorithms has increased motivated by the attainment of high accuracies, the side effect has been a decrease in the interpretability of models. Little emphasis has been placed on clarifying the internal mechanics of the models. Examples of studies eliciting explanations of the models' reasoning underpinning their forecasts do not yet exist. Traditional time-series and simple regression methods have produced reasonable accuracies in previous studies. These approaches are still used and they continue to produce competitive results for forecasting patient flows. Research has shown that these techniques have excelled particularly in cases when the underlying data possessed consistent variations; however, it has been noticed that their accuracies become compromised when sporadic fluctuations are encountered [21]. Approaches that combine the advan-

tages of traditional approaches with those of machine learning have not yet emerged in this domain.

The features used for the forecasting models in the prior works have chiefly relied on autoregressive (previous or lagged values) and various calendar variables [5, 26, 48] with some studies using variables indicating school holidays as well. There has been a recent trend to also incorporate weather-related variables [7, 37, 42, 53], while air quality and pollution information have been a subject of some studies [32, 40].

The aim of this study is to present a machine learning solution to forecasting total daily patient flows 7-days ahead in order to facilitate resource planning strategies at UCC facilities. This work particularly aims to investigate the feasibility of a range of real-time proxy variables to contribute to improving the overall model accuracies, as well as to employ a range of eXplainable Artificial Intelligence (XAI) techniques to both bring interpretability into the model behaviour and to extract additional insights. Additionally, this research seeks to develop mechanisms for enabling the models to adapt to the disruptive conditions of the recent COVID-19 pandemic, with learnings that can be applied to similar scenarios in future.

1.1 Contribution

The principal contribution of this study lies in the integration of novel, near real-time[1] proxy features aimed at maintaining the precision of patient-flow forecasts during periods of instability, such as the COVID-19 pandemic. The value of these proxy variables within the broader context of concept drift is highlighted as a crucial takeaway for practitioners. Additionally, the study also makes secondary contributions through the utilisation of previously unexplored machine learning models and the large number of algorithms used in experiments. A further contribution of this study is the combination of traditional and machine learning approaches in order to leverage the advantages of both. Furthermore, the study also presents the introduction of XAI techniques for the inspection of models' behaviour in this domain, providing an understanding of both the high-level characteristics of the models and their reasoning for individual forecasts.

2 Related work

The vast majority of prior research in forecasting patient flows has focused on EDs. Prior studies [10, 13, 48] have used patient flow forecasting and analysis literature from EDs and UCCs interchangeably and thus we follow the same pattern with the understanding that approaches that have been effective in the ED domain are transferable to the UCC context, and vice versa since the underlying mechanisms of demand overlap. Indeed, in their systematic literature review, Coster et al. [10] identify six key drivers of patient flow demand common to both EDs and UCCs. These included limited access to or confidence in primary care, patient-perceived urgency, convenience, recommendations by family and friends, or other healthcare professionals, and the belief that their condition required the resources and facilities offered by a particular healthcare provider. While theoretically valuable, none of these variables are generally available for use in forecasting patient demand. He et al. [22] also add to this list, population growth and ageing which carry a

---

[1]The term real-time has different meanings depending on a domain. Real-time in this context is defined by data that has been updated within a 24-hour window.

Table 1 Summary of literature predicting total daily patient flows in EDs, highlighting forecasting horizons, best algorithms and the accuracies achieved

Study	Year	Forecast periods (days)	Features	Best algorithm	MAPE
Boyle et al. [5]	2012	1	4	ARIMA, ES, OLS	7.0%
Xu et al. [51]	2013	1	7	ANN	6.8%–7.3%
Marcilio et al. [29]	2013	7	15	GLM	7.6%
		30		GLM	9.7%
Xu et al. [52]	2016	1	31	ARIMA-LR	6.5%
		7		ARIMA-LR	9.6%
Calegari et al. [7]	2016	1	5	SES	2.9%
		7		SES	10.7%
		14		SES	10.7%
		21		SES	11.4%
		30		SES	11.7%
Navares et al. [32]	2018	1	13	ARIMA	8.1%–12.3%
Whitt and Zhang [49]	2019	1	57	SARIMAX	8.4%
Rocha and Rodrigues [39]	2021	1	15	LSTM	4.2%
Sudarshan et al. [42]	2021	3	10	CNN	9.2%
		7		LSTM	8.9%
Vollmer et al. [47]	2021	1	30	Stacking	6.8%–8.6%
Harrou et al. [21]	2022	1	7	DBN	4.1%
Zhang et al. [53]	2022	1	29	SVR	8.8%
Petsis et al. [37]	2022	1	38	XGBoost	6.5%
		2		XGBoost	6.9%

greater utility for forecasting long-term patient demand trends. More general causal factors for UCCs patient demand are discussed by DeLurgio et al. [13], where the authors list increasing trends in certain diseases like the prevalence of diabetes, seasonal influences like holidays as well as cyclical and irregular factors like flu epidemics and natural disasters, which also broadly align with those EDs [16].

We review published literature which is most comparable to our work in terms of studies that have predominately investigated the forecasting of total daily volumes of patient arrivals. We also focused on studies which reported their forecasting accuracies in terms of the Mean Absolute Percentage Error (defined in Equation (1)) which to some degree enables comparisons across different studies. Forecast horizons in terms of the number of days ahead being estimated are relevant for this study, and these are also highlighted and summarised where possible in Table 1.

Traditional regression and autoregressive modelling dominated the earliest research into forecasting patient demand at EDs. Batal et al. [2] successfully used multiple linear regression (MLR) to predict patient flows by incorporating a selection of calendar variables like the day of the week, month, season, and holiday flags. Jones et al. [26] subsequently used the same set of variables together with MLR as a benchmark model against comparisons with autoregressive ARIMA and SARIMA models for predicting daily patient presentations. Their work fundamentally differed in that they extended their forecast horizon to 30 days ahead. Meanwhile, Boyle et al. [5] focused on next-day ED patient flow forecasting, while using MLR, ARIMA and Exponential Smoothing (ES) models. In contrast, both Champion et al. [9] and Aboagye-Sarfo et al. [1] altered the frequency from daily, to predicting the total monthly patient demand. Both studies relied on ES with ARIMA modelling methods, while Aboagye-Sarfo et al. [1] also integrated these algorithms with experiments using VARMA.

The focus of subsequent works explored the viability of forecasting at differing frequencies together with longer forecasting horizons. One-step-ahead hourly, daily and monthly ED patient flow forecasting models with the use of calendar information (day of week, month and holidays) were developed by Boyle et al. [5]. Their work also experimented with standard approaches like ES, ARIMA and MLR, concluding that the errors increased and thus the forecasting became less reliable as the forecast granularity became finer. Marcilio et al. [29] returned to forecasting daily patient flows, exploring 7 and 30-day ahead predictions with the help of calendar as well as climatic variables. They expanded the range of techniques used to Generalized Linear Models (GLM), Generalized Estimating Equations (GEE) and Seasonal ARIMA (SARIMA). Whilst recognising the importance of forecasting further out into the future, Calegari et al. [7] addressed predicting total daily ED patient flows over 1, 7, 14, 21 and 30-days ahead, while also relying on the calendar (day of week, month and holidays) and climatic data. They also used standard SARIMA as in previous studies, but also included Seasonal ES (SES), Seasonal Multivariate Holt Winter's ES (HWES) and Multivariate SARIMA. Xu et al. [52] proposed a method that combined ARIMA and Linear Regression (ARIMA-LR) to predict daily ED patient flows for both the next day as well as 7-days ahead forecasting. Given the effectiveness of traditional approaches in this domain, more recent studies have continued to leverage these techniques like Carvalho-Silva et al. [8], who experimented with forecasting at different levels of granularity. They forecasted total patient flows on a next-week and next-month basis using ARIMA and ES, while Whitt and Zhang [49] forecasted next-day total patient flows using SARIMAX using also calendar and climatic variables.

Though traditional time-series and regression approaches possess more assumptions and stricter requirements that are harder to satisfy, they do continue to generate effective solutions for forecasting patient flows. In contrast, non-parametric machine learning approaches avoid many of the rigid assumptions required by traditional statistical methods and therefore have some added flexibility accompanied by a proclivity to overfit. Over the past decade machine learning solutions in this domain have started to emerge and have consistently produced competitive accuracies, while also exceeding those of traditional approaches in some cases [39, 51]. Machine learning methods generally have the ability to capture non-linear relationships in the data and thus hold the potential to produce models which arguably better represent the complex and dynamic nature of patterns in this domain [53].

Xu et al. [51] was one of the earliest works using machine learning in this domain where they applied Artificial Neural Networks (ANN) in conjunction with variables describing seasonal influenza epidemics, as well as calendar and climatic data in order to forecast one-day ahead total daily patient flows. We see the emergence of ensemble-based techniques like Random Forest (RF) and Gradient Boosting Machines (GBM), together with ANN for forecasting daily respiratory and circulatory-related ED admissions in work by Navares et al. [32]. In this study, the authors incorporated environmental and bio-meteorological variables into the models that captured air quality indicators, and ultimately concluded that machine learning methods did outperform ARIMA when the models were combined using Stacking.

We are now beginning to see an emergence of a broader range of proxy variables being used, together with deep learning approaches and some early signs of XAI techniques entering this domain. Vollmer et al. [47] proposed patient flow models forecasting 1, 3

and 7-days ahead while also using RF and GBM, together with a k-Nearest Neighbours (kNN) regressor and stacked generalisation combining machine learning and statistical approaches. They expanded the range of proxy variables used in their models and included weather, school and public holidays, and seasonal pattern data, alongside large scheduled events which tended to see an influx of patients. Additionally, they also utilised Google search data for the keyword "flu". Long Short-Term Memory (LSTM) was used by Rocha and Rodrigues [39] where they performed a wide range of comparative experiments with those of traditional ES and SARIMA approaches and other machine techniques like Autoregressive Neural Networks (AR-NN) and XGBoost (incorporating calendar variables like day of week, month and holidays). The authors concluded that LSTM produced the best accuracies for next-day forecasts of total daily patient flows. Sudarshan et al. [42] likewise used LSTM for predicting 3 and 7-day ahead total daily patient flows. They included Convolutional Neural Networks (CNN) into their suite of algorithms and compared the results against RF. The authors also confirmed the effectiveness of deep learning methods in this domain by demonstrating that LSTM was most accurate for the 7-day forecasting, while CNN was more suited for 3-day ahead forecasting in their setting.

Harrou et al. [21] expanded the range of deep learning techniques in their work by including Deep Belief Networks (DBN), Restricted Boltzmann machines (RBM), Gated Recurrent Unit (GRU), a combination of GRU and Convolutional Neural Networks (CNN-GRU), CNN-LSTM, as well as the Generative Adversarial Network based on Recurrent Neural Networks (GAN-RNN). They designed their models to predict total daily patient flows one-day ahead, finding that the best accuracy was achieved using DBN. However, research also demonstrated that deep learning methods do not always deliver the best accuracies.

The study by Zhang et al. [53], contrasted LSTM with standard machine learning methods like kNN, Support Vector Regression (SVR), XGBoost, RF, AdaBoost, GBM and Bagging, in addition to traditional methods like OLS and Ridge Regression. This study found that it was SVR which outperformed all other algorithms based on models predicting next-day patient flows while using calendar and meteorological variables.

Ensemble-based techniques leveraging GBM such as XGBoost and LightGBM continue to increase in prominence. Guyeux and Bahi [20] used both algorithms together with RF and Lasso for hourly forecasting of patient flows. Similar to our study, the authors expanded the use of variables for forecasting beyond using standard calendar and meteorological variables. Instead, the study augmented these variables with road traffic information and epidemiological data reporting the current number of cases of influenza, chicken pox and acute diarrhoea. All in all, the study used 712 explanatory variables for forecasting and concluded that XGBoost was best performing in their setting. Meanwhile, another recent study bearing parallels with our research was conducted by Petsis et al. [37], who focused on predicting daily patient flows 1 and 2 time periods ahead. This work is reported to be the first study that has started to incorporate XAI techniques into illuminating the behaviour of the underlying models. This work primarily considered the high-level interpretability of the models which included the analysis of feature interactions, but the reasoning behind the forecasts for specific days was not considered. The study again used Gradient Boosting in the form of XGboost, relying on standard calendar and weather variables which amounted to a total of 38 variables.

2.1 Summary

It is natural to see that traditional techniques for forecasting patient flows were the most prominent approaches in earlier research. The emerging trend within the existing literature points in the direction of increasing use of machine learning algorithms, particularly those with more inherent complexity. Though machine learning approaches are achieving a widespread uptake, traditional methods continue to be used as well and generate competitive accuracies; however, they are now more often used as benchmarking models. This can be explained by the fact that machine learning algorithms tend to be a preferable choice over traditional statistical methods in forecasting, as they can handle large and complex data, adapt to changing environments and more easily handle missing and high dimensionality data which can be seen in features increasing over time (Fig. 1a).[2]

Thus far, the literature indicates that amongst the researchers autoregressive, calendar and holiday variables, together with weather and bio-meteorological features remain popular model inputs for forecasting. Some recent investigations into the usefulness of the proxy variables like internet search terms and live influenza tracking indicators are emerging. Due to this and an increasing number of features being used, the suitability of non-parametric algorithms becomes evident. Traditional approaches possess advantages which can still be leveraged. However, studies predominately focus on exploring either traditional or machine learning approaches instead of attempting to combine both into a single model that is able to leverage the advantages that both techniques offer. To that end, this study seeks to fill this gap. Additionally, the literature shows that there is an emergence of an ever more diverse range of machine algorithms being used in this domain; however, there is no clear algorithm which stands out as the best-performing.

Achieving high accuracy in forecasting patient flows has been a central focus of most prior studies. Figure 1b depicts the trend in the reported forecasting accuracy over time. The figure reveals a gentle downward slope indicating a gradual reduction in error (reported as the Mean Absolute Percentage Error—MAPE) values over time.[3] It should be noted that only a single study used datasets covering the COVID-19 period which exhibited sudden and recurring concept drifts in the underlying data patterns, which result in deteriorations of accuracy. Given this gap, this study aims to explore the effectiveness of a larger set of novel proxy variables to mitigate the negative effects of concept drifts, such as the most recent one during the pandemic period.
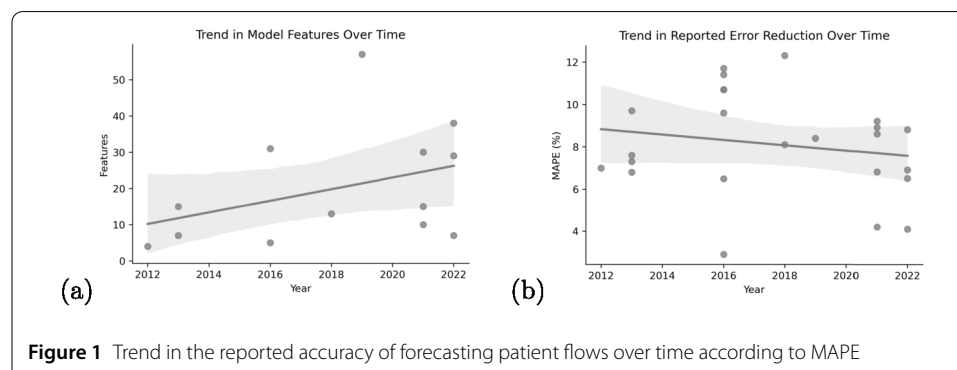

(a)   (b)

Figure 1 Trend in the reported accuracy of forecasting patient flows over time according to MAPE

---

[2]The trend lines do not represent a statistically significant line of fit.

[3]The overall picture presented in the figure is arguably optimistic until newer studies emerge and are published which include results from this period.

Finally, the increased usage of more complex machine learning algorithms is accompanied by the generation of uninterpretable "black box" models. A clear gap in the literature for patient flow forecasting exists in the absence of analyses which consider both the high-level mechanics of model behaviour as well as the interrogation of the models' reasoning into specific forecasts. No study as yet using machine learning has explored the internals of the models to the fullest extent, which this study attempts to address.

In view of the existing literature, the research questions (RQ) addressed in this study are:

- (RQ1) Which machine learning models attain the highest improvements over the benchmark strategies for short-term forecasting of daily patient arrivals at UCCs? Can machine learning and traditional statistical approaches be combined to produce higher accuracies?
- (RQ2) Can proxy variables improve forecast accuracies and adapt the models to the concept drift caused by the COVID-19 conditions? Which variables are most effective at generally improving the forecasting accuracies of daily patient demand seven days ahead?
- (RQ3) How can greater interpretability and explainability of forecasting models be achieved?

3 Methodology

3.1 Setting

The data were sourced from Shorecare[4] which owns and operates the clinics in this study located in Auckland, New Zealand. The Smales Farm clinic provides 24-hour care services, while the Northcross clinic is reduced to after-hours care. The clinics treat standard low acuity cases and provide x-ray and fracture clinics as well as facilities for complex wound management. The Smales Farm clinic is the sole 24-hour UCC servicing a population of approximately one-quarter of a million and is situated within one kilometre of a major hospital whose ED treats ~46,000 patients annually.

3.2 Patient flow dataset

Models were designed for predicting daily arrivals seven days ahead. The dataset recorded patient presentations spanning 11 years from 2011 through to 2022. Figure 2 shows the characteristics of the patient flows on a selection of the dataset for both clinics. Seasonal patterns indicating increases in patient arrivals are visible during the winter months of the Southern Hemisphere (June-August). This can be explained by increased presentations of Influenza and other respiratory-related illnesses. The beginning of the period affected by strong concept drifts is highlighted and deviations from the prior patterns can be observed. Additionally, the exact points corresponding to the most stringent mandates concerning COVID-19 pandemic lockdowns are highlighted, as well as periodic partial closures of the smaller clinic during portions of this period.

An aggregation of the data by week number across all years can be seen in Fig. 3. There is generally a reduction in demand during the school holidays, with peaks in patient arrivals coinciding with the Winter months (mid-year), and the end of the year when the local general practitioners (GPs) are closed for holidays.

---

[4]https://www.shorecare.co.nz/

Figure 2 Daily patient demand at both UCC clinics ranging from 2015–2021

Figure 3 Average patient demand across both UCCs for all years by week number
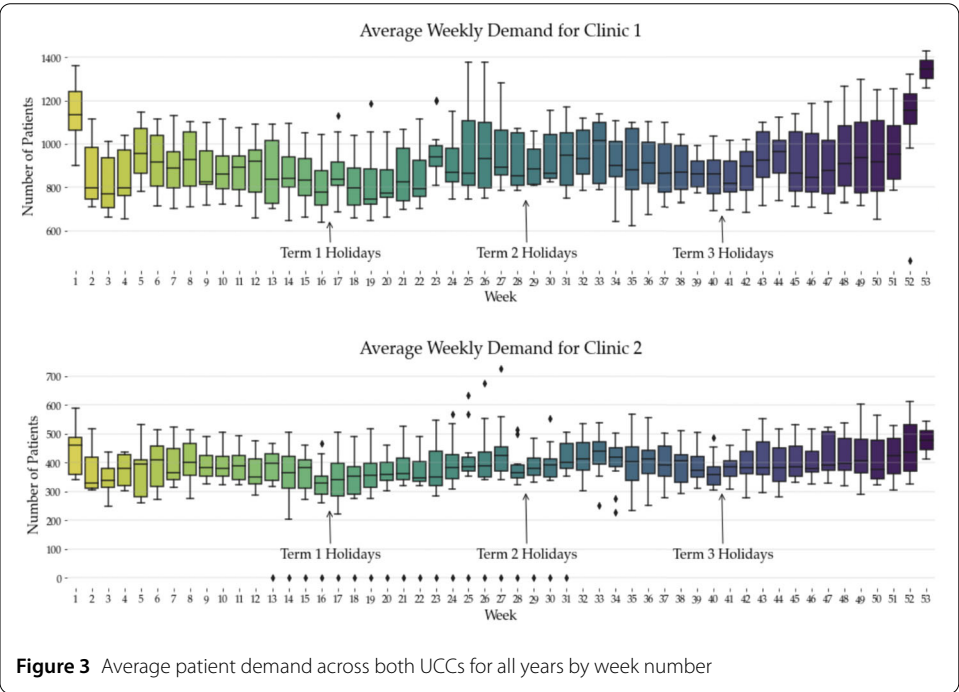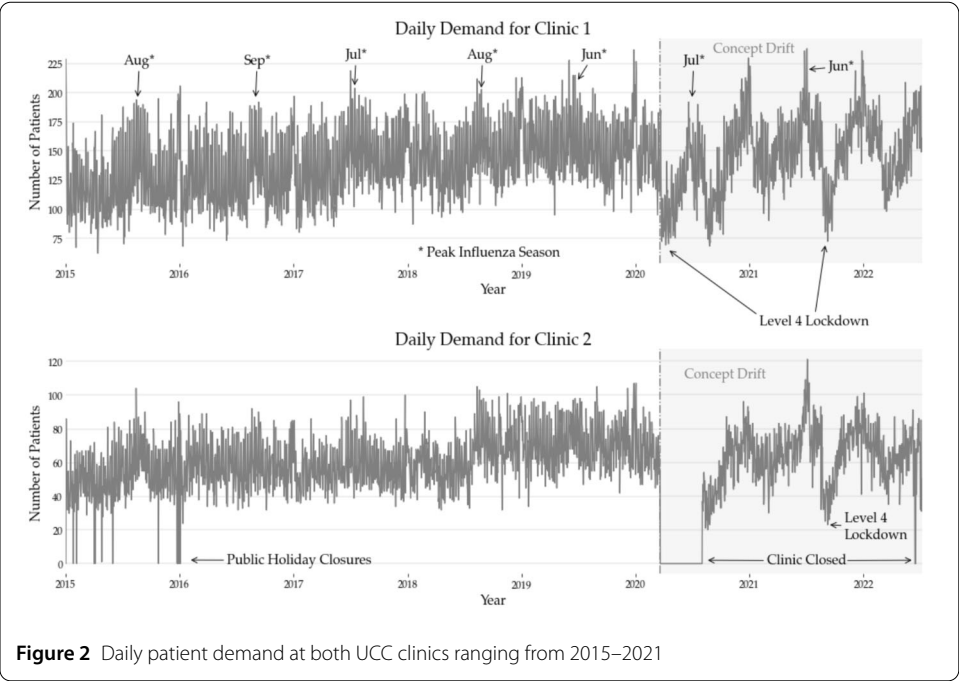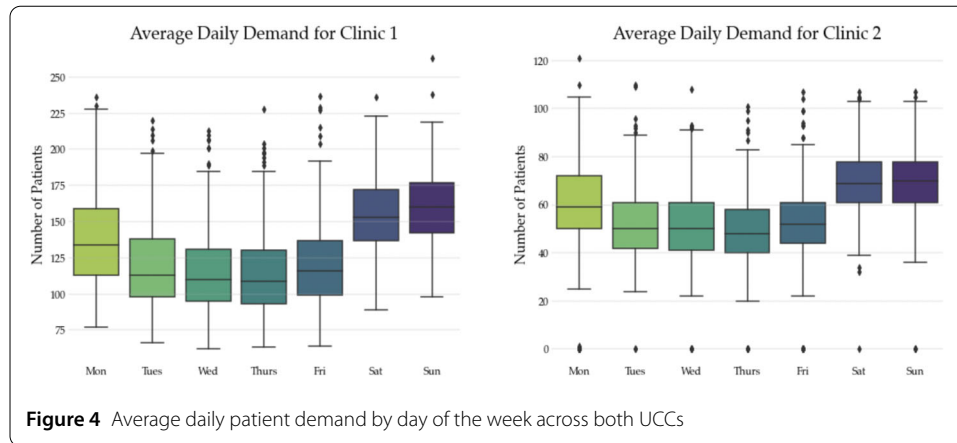
Figure 4 shows average patient arrival patterns across both clinics by day of the week. A peak in patient presentations tends to occur during the weekend when GPs are closed, which is followed by a decrease in patient flows at the start of the working week, which then reaches minima by mid-week.

Figure 4 Average daily patient demand by day of the week across both UCCs

3.3 Proxy data and features

The models use a mixture of autoregressive and proxy features. Since the data exhibits pronounced seasonal and weekly cyclical patterns, we used autoregressive features representing previous patient arrival values from 7, 14, 364, 728 and 1092 days before as predictors. Additionally, the week number in the year was used to provide information describing the seasonal trends as well as recurring school holiday patterns. A public holiday flag was devised which ensured that the elevated demand driven by these days was represented.

A strategy was devised to handle the pandemic-related concept drift by using a set of variables which had the ability to capture the changes in the underlying conditions. Primarily, we created a variable that represented the severity of the legally mandated COVID-19 restrictions[5] prevailing within the clinics' locality.

Other variables were used with the assumption that they would capture both the signal indicating that concept drift was occurring, and that they would also contribute to the general improvement of the overall models by being effective proxies for patient demand. In order to qualify for consideration and be useful for real-time forecasting, these variables had to satisfy three constraints. The first being that the data needed to be measured in real-time and thus be up-to-date, secondly, the data had to be publicly available in real-time, and finally a sufficient amount of historic values of this data had to be available for model training.

Google Trends [18] data provides information on the frequency of Google search keywords over a given time period and region and satisfies the above constraints. We used this data as a proxy variable with the assumption that it could capture potential increases in various symptoms which may trigger higher patient arrivals. The selected keywords were *flu, headache, sick, chest pain* and *cough* being both generic as well as relevant to the COVID-19 pandemic. Google Trends data is accessible at weekly aggregate results provided that the search criteria fall within 5 years.

We considered the possibility of including pedestrian foot traffic as an additional proxy variable in our models, based on the underlying assumption that higher levels of pedestrian activity would indicate higher public activity levels and, therefore, an increase in

---

[5]This expressed either the COVID-19 Alert Level system [33] or subsequent Traffic Light mappings [34] defined by the New Zealand Government which were later introduced.

Table 2 Feature names as they appear in figures and their descriptions

Feature name	Description
lag7d	Autoregressive 7-day lag, the value from one week prior
lag14d	Autoregressive 14-day lag, the value from two weeks prior
lag1	Autoregressive 364-day lag, the value from one year prior
lag2	Autoregressive 728-day lag, the value from two years prior
lag3	Autoregressive 1092-day lag, the value from three years prior
public_holiday	Public Holiday Indicator (0/1)
week	Week Number ranging from 1–53
ped_count	Auckland CBD foot traffic
flu_percent	weekly reports on Influenza prevalence in New Zealand
covid_level	1–4. Prior to 3/12/2021: COVID-19 Alert Level ranging from 1–4; After 3/12/2021: Traffic Light Mappings (Green = 1; Orange/Red = 2; No mappings to 3 or 4)
trends	Google Trends normalised frequency of term searches of *flu*, *headache*, *sick*, *chest pain* and *cough* as a single signal
feels_like	A combination of temperature, wind chill and heat index values

patient arrivals. For this, we used data from Heart of the City Auckland [23], which provides information on the number of pedestrians walking past approximately 20 cameras in the Central Business District (CBD) of Auckland each day, with data stretching back to 2012.

Similar to previous studies, we also integrated information on the prevailing weather conditions. We sourced the data using APIs from Visual Crossings [46]. Our exploratory analysis indicated that the most effective feature from the suite of possible weather variables was the *'feels like'* indicator that combines temperature, wind chill and heat index values.

Guided by the advice of domain knowledge of the UCC staff, we also incorporated the data from FluTracker [17]. FluTracker monitors prevailing levels of Influenza in New Zealand using surveys. Approximately 30,000 people report weekly symptoms of fever or cough being experienced. The New Zealand FluTracker data is available only from 2018 onwards, with weekly reports provided online. The data also has missing values which coincide with the summer months when the prevalence of the illness is usually low. The data is presented as the percentage of respondents exhibiting symptoms each week.

Table 2 lists the names of all the variables used in various figures together with their descriptions. Appendix B shows end example of the proxy feature values versus patient flows in Fig. 11 to Fig. 14.

3.4 Benchmark models

We evaluated the efficacy of the proposed models against several benchmark models. The primary benchmark model replicated the current in-house strategy employed by the clinics' administrators to estimate patient arrivals. This approach used the patient arrival numbers from the same period in the previous year plus an additional 5% to account for an increasing trend in total volumes.

The second benchmark model is a Persistence Model. It is essentially a Random Walk technique [35] which makes the forecasted value the same as that of the identical day in the previous year. The third benchmark model was ARIMA [4], created with auto-tuning. The final benchmark model consisted of an enhanced version of the Persistence Model which forecasted a value for a given day at a point in time t to be a mean of weighted autoregressive features consisting of values from t-7, t-14, t-364, t-728 and t-1092 days.

3.5 Algorithms

We used ten statistical and machine learning algorithms in order to generate candidate models. These consisted of: Random Forest (RF) [6], Voting [36], Stacking [50], Ridge Regression [24], Support Vector Machines Regression (SVR) [15], Kernel Ridge Regression (KRR) [12], K-Nearest Neighbour Regression (kNN) [11] as implemented in Scikit-learn [36], CatBoost [38], Prophet [44] and an Averaging Model. Table 3 lists all the algorithms as well as the benchmark models with their brief descriptions, together with their hyper-parameter values where relevant.

3.6 Testing approach

The models were tested on data covering five and a half years from 2017 through to mid-2022. A modified version of the expanding window approach was used for estimating the generalisability of the models. Up to a maximum of 5 previous years of training data were used for creating each model.[6] Following each training phase, the models were then tested on forecasting 7 days ahead. There were in total of 286 test sets in this hold-out approach. The models were initially trained on data from 2014–2016[7] and the forecasts were then made starting from 1 January 2017 up to 7 days ahead. Figure 5 visually depicts our training/testing methodology. The whole testing process was performed three times; once for models containing all the proposed features. The second time for models using only autoregressive features to establish the efficacy of the proxy features. Finally, the models were trained using only proxy features in order to both assess their utility and extract additional insights.

3.7 Model error measurements

The models were evaluated using several metrics with each one providing a slightly different perspective.

Mean Absolute Percentage Error (MAPE) is frequently used in literature and is recommended as the primary evaluation metric for forecasts [3]. We therefore followed this recommendation. MAPE is scale-independent and can be used to compare forecasts across datasets and studies with different ranges of values for the dependent variable. The calculation of MAPE is as follows:

$$\text{MAPE} = \frac{100}{T} \sum_{t=1}^{T} \left| \frac{\hat{y}_t - y_t}{y_t} \right|, \tag{1}$$

where T is the number of forecasts under evaluation and $\hat{y}_t - y_t$ is the error or residual term arising from the difference between the observed y value and the forecast value $\hat{y}_t$ at time point t. MAPE is a useful measure because it is able to express deviations between the observation and the forecasted values in terms of percentages, and as such, it is easy to interpret.

---

[6]The models were trained on a maximum of five years of historic data in part due to the preference of more recent data for modelling, and due to a limitation in the acquisition of the Google Trends search terms data. Google Trends data can only be extracted for up to five years at a weekly granularity and beyond this, the data is aggregated at a monthly granularity. Since Google Trends data is scaled and provided as a relative index, it made it unsuitable to concatenate different periods together.

[7]Data prior to 2014 could not be used for training due to the autoregressive values which created missing values in the initial few years.
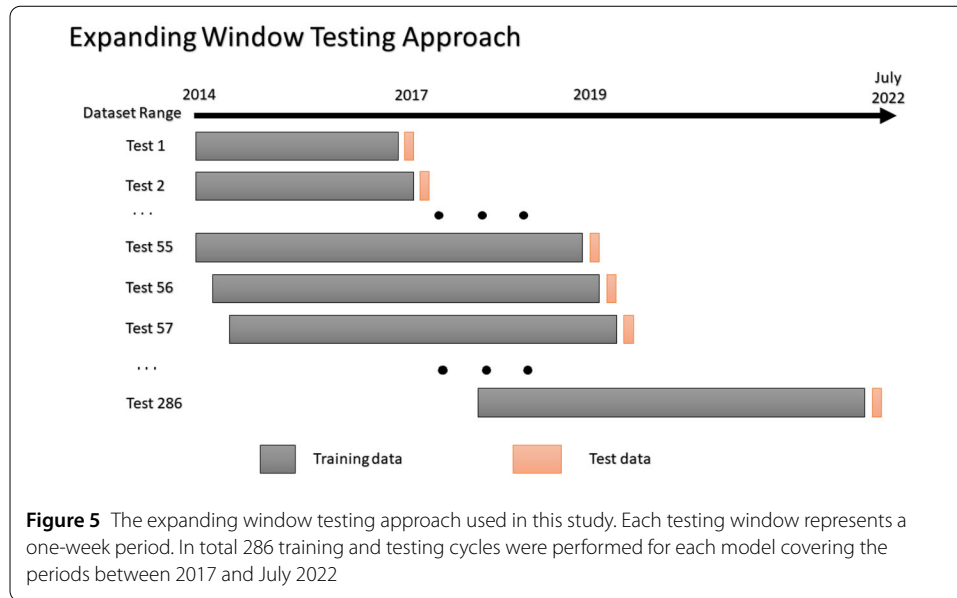
Table 3 Summary of the benchmark models and algorithms used in this study, together with hyperparameter settings where applicable

Method	Description
Benchmark	Current estimation method used in-house by the clinics which forecasts patient demand for a given day to be 5% higher than that of the same day in the previous year.
Persistence Model	A benchmark model implemented as a Random Walk [35] method with the forecast being the same as the value for the same period of the previous year
Enhanced Persistence Model	An optimised benchmark model that made forecasts based on the weighted mean value of autoregressive values in respect to time t with time lags of t-7, t-14, t-364, t-728 and t-1092. The weightings were optimised through an empirical approach and set as [5, 4, 3, 2, 1] respectively for each autoregressive feature, with features representing recency being allocated greater importance.
ARIMA	Traditional autoregressive statistical technique, predicting future values based on past values.
kNN Regression	A non-parametric algorithm that bases its predictions on the principle of proximity, producing a forecast that is an aggregation of k nearest observations with respect to the characteristics of the data point in question.
Ridge Regression	A technique that creates a parsimonious model which shrinks the coefficients towards zero using L2 regularization. The resulting models generally reduce the variance resulting in an improved mean-squared error.
Support Vector Machines Regression	SVR is an extension of SVMs. It uses a pre-defined kernel function to transform the data from a non-linear space to a higher dimension in order to find an approximate fit that satisfies a pre-determined error margin. To that end, the objective function of SVR is to reduce the coefficients rather than the error term (epsilon). SVRs are particularly effective on smaller datasets and are more robust to outliers.
Kernel Ridge Regression	Kernel ridge regression extends Ridge Regression with the integration of the kernel trick technique from SVR. It differs to SVR in that it uses the squared error loss as opposed to the epsilon-insensitive loss in SVR, combined with l2 regularization.
Prophet	Auto-tunable, additive forecasting model with the ability to handle non-linear trends using yearly, weekly, and daily seasonality with capabilities to integrate effects from holidays, having robustness to dislocations in trend.
Random Forest Regression	Ensemble-based algorithm consisting of decision trees whose outputs are combined. Each decision tree is induced based on random feature subsets, resulting in an uncorrelated forest of trees. The combined accuracy of the forest results in a higher fidelity than that of any individual tree.
CatBoost	CatBoost is an ensemble-based algorithm that generates gradient-boosted decision trees. During training, successive trees are induced with a reduction in loss. The size of the ensemble is preset by defining the maximum number of trees as a parameter.
Voting Regressor	Ensemble-based meta-estimator. Combines machine learning and traditional time-series approaches. Initially generates models for the underlying base regressors: Prophet, CatBoost, Random Forest and ARIMA. It then combines the outputs of these algorithms for the final forecast using a weighted combination scheme.
Averaging Model	This algorithm was a customised version of the Voting Regressor which combined the outputs of five algorithms (Prophet, CatBoost, Random Forest, Voting and Stacking) but discarded the highest and lowest predictions in the calculation of each prediction.
Stacking	An ensemble-based meta-estimator which models the forecast outputs of the underlying base estimators (Prophet, CatBoost and Random Forest) using an overarching regressor whose output constitutes the final forecast.

We also note the Root Mean Square Error (RMSE) for each model as defined below:

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}. \tag{2}$$

RMSE is instructive since it describes the dispersion of the errors while being scaled to the dependent variable, therefore smaller RMSE values are preferred.

Figure 5 The expanding window testing approach used in this study. Each testing window represents a one-week period. In total 286 training and testing cycles were performed for each model covering the periods between 2017 and July 2022

For completeness, we also report the Mean Absolute Error for all models, which in conjunction with the previous two metrics is also used by some more recent studies in this field [53]. MAE is the average absolute difference between the observation and the forecasted values. Given this property, it is to some degree conceptually easier to interpret and due to the squaring of the differences, it is less sensitive to large errors, unlike RMSE. Therefore, several significant errors will influence RMSE to a larger extent than MAE. The calculation for MAE is:

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |\hat{y}_t - y_t|. \tag{3}$$

Lastly, we extensively use mean ranks in order to concisely summarise the performance of all the algorithms across every test dataset. For each forecast period of 7-days, every algorithm was ranked from 1 to 15 with respect to its MAPE value, with the best performing algorithm achieving the rank of 1. This was performed across all 52 testing periods per year and across each of the 5.5 years of testing, and from this, the mean ranking was calculated.

3.8 Statistical measurements

In addition to the evaluation metrics, we also use Theil's U statistic [45] in order to assess the model accuracy relative to the persistence model where the forecast value is equal to the previous value. Since the method squares the errors, it gives more weight to large deviations and exaggerated them, which can serve as a useful method for identifying suboptimal models. The Theil's U statistic value is calculated below as:

$$U = \sqrt{\frac{\frac{1}{T} \sum_{t=1}^{T-1} \left(\frac{\hat{y}_{t+1} - y_{t+1}}{y_t}\right)^2}{\frac{1}{T} \sum_{t=1}^{T-1} \left(\frac{y_{t+1} - y_t}{y_t}\right)^2}}, \tag{4}$$

where y again is the observed value and $\hat{y}_t$ is the forecast value at a given time step t. When interpreting this statistic, values lower than 1 indicate that the model is performing better than the persistence model, while values of 1 and beyond indicate that the forecast accuracy is equivalent to the persistence model and is in fact worse as the values increase.

Finally, we report the Diebold and Mariano [14] statistical test to establish whether the sequences of forecasts of the models are meaningfully different from those of the benchmark. In order to determine this, we use this test to compare the outputs of the competing model with those of the benchmark estimation models, with respect to the observed values.

3.9 Model interpretability and explainability

The emerging field of XAI, spurred on by increasing regulatory requirements [30], addresses the challenge posed by uninterpretable models and attempts to answer both the "how" and the "why" of their decision-making in approximate terms. A set of approaches called post-modelling explainability tools, aim to answer "how" an algorithm behaves in the construction of a model during the training process, resulting in model interpretability, as well as "why" the generated model has made a specific prediction/forecast, resulting in model explainability [31]. One of the techniques that currently stands out as state-of-the-art for extracting the interpretability and explainability of predictive models [19], is SHAP [27], which is used in this study.

We employ SHAP to examine the internal mechanics of the predictive models at both the *global* and *local* levels. At a global level of analysis, we seek to understand the overall effects that each feature exerts on the model outputs. We primarily use feature importance plots to gain this insight which typically ranks as well as depicts the relative impacts of each feature. We also examine the effects that changing feature values have on the final forecast. Additionally we use feature dependence plots in order to shed light on how pairs of features interact in order to affect the final forecast. These plots together offer a degree of *high-level* interpretability of the main drivers for a given model. In considering model behaviour at a local level, we attempt to extract a model's reasoning as to precisely why exactly it has produced a given forecast for a specific data point.

SHAP generates new models which approximate the forecasting behaviour of the underlying "black-box" models. These models are called *surrogate models* and are designed to be more interpretable. SHAP (an abbreviation for SHapley Additive exPlanation) itself is based on Shapley values [41]. The technique operates on the principles of game theory. It attributes each feature's marginal contribution to the final predictive outcome in collaboration with the other features. In this way, SHAP is able to provide both *global* interpretability and *local* explainability.

4 Results

The results are presented in two parts. The first examines the accuracies and the statistical significance of the forecasting models, together with the efficacy of the proxy features. The effects of the concept drift on the model accuracies are highlighted from the year 2020 onward. The second part analyses both the interpretability and the explainability aspects of the models, with a focus on determining the utility of the proxy variables. The second part also attempts to extract insights concerning the model behaviour and the underlying features.

4.1 Forecast modelling

Table 4 shows a high-level summary of all the models across both clinics, displaying the MAPE values for each candidate model developed with a full set of features. To establish the utility of the proxy features, each model's MAPE score is contrasted with the MAPE values of models developed using only autoregressive features, which are placed adjacently in parentheses. Both the RMSE and MAE values are also listed, together with the average rank scores based on MAPE. The table indicates that across both clinics the proposed models have outperformed those that have not used proxy features. At the bottom of the table, it can be seen that the best-performing algorithm was Voting which combined machine learning and standard statistical approaches.

The table combines the rank-ordering of the modelling techniques from worst performing to most accurate with respect to MAPE across both clinics (seen in the final column of Table 4). The results indicate that the Voting algorithm has consistently outperformed all other techniques on this dataset by leveraging the advantages of both machine learning and statistical approaches. The top three performing techniques are all ensemble-based meta approaches which have in various ways combined the models from the underlying algorithms. These approaches were closely followed by standard ensemble-based methods, with Gradient Boosting-based approaches outperforming Random Forest. The rank-ordered results also indicate that the existing in-house benchmark approach to patient forecasting has been significantly improved upon by the best-performing methods.

It is also observable through MAPE scores that the predictability of Clinic 2 patient flows is generally lower than that of Clinic 1. This is attributable to overall lower patient flows at Clinic 2 which predisposes it to more variability and thus higher unpredictability. This result underscores the limitations of blindly using the MAPE measure for comparisons across different studies without taking volumes into consideration since the magnitude of the total patient volumes affects variability and consequently, predictability. In saying that, comparative studies from reported literature cite MAPE accuracies ranging from 7.6% to 10.7% for 7-day forecasts (Table 1) which provides some context for the 9% and

Table 4 Forecast accuracies by algorithm and clinic using all features as well as MAPE accuracies of models using only autoregressive features in parentheses. Ranks per clinic are based on MAPE and are also combined across both

	Clinic 1				Clinic 2				Combined
	MAPE	Rank	RMSE	MAE	MAPE	Rank	RMSE	MAE	Rank
kNN	17.5 ± 8.8 (11.3)	12.7	29.8	25.2	21.7 ± 14.7 (15.0)	12.3	15.7	13.2	12.5
SVR (NU)	17.7 ± 11.3 (11.2)	12.4	29.7	25.3	20.4 ± 16.8 (14.5)	11.1	14.6	12.4	11.7
Benchmark	16.7 ± 13.5 (16.7)	11.4	25.6	22.0	21.9 ± 21.1 (21.9)	11.1	14.6	12.4	11.2
Naive	16.3 ± 11.7 (16.3)	11.0	25.7	22.3	19.1 ± 16.6 (19.1)	10.3	13.5	11.4	10.6
Prophet	10.1 ± 5.3 (11.8)	6.9	16.7	14.2	15.7 ± 14.6 (16.6)	7.9	10.9	9.2	7.4
Naive (Enhanced)	10.7 ± 6.1 (10.7)	7.6	17.7	14.9	14.1 ± 9.4 (14.1)	7.0	10.1	8.4	7.3
ARIMA	10.5 ± 5.9 (10.5)	7.4	17.3	14.6	14.4 ± 9.6 (14.1)	7.2	10.1	8.5	7.3
Kernel Ridge	9.9 ± 4.2 (10.6)	7.0	16.8	14.1	14.1 ± 9.0 (14.3)	7.5	10.1	8.3	7.3
Ridge	9.9 ± 4.2 (10.6)	6.9	16.8	14.1	14.1 ± 9.2 (14.3)	7.3	10.1	8.3	7.1
Random Forest	10.0 ± 5.6 (10.5)	7.1	16.7	13.9	13.9 ± 9.1 (14.2)	7.1	10.0	8.3	7.1
CatBoost	9.6 ± 4.4 (10.4)	6.8	16.3	13.5	13.4 ± 7.0 (14.2)	7.1	9.8	8.1	7.0
Gradient Boosting	9.6 ± 4.5 (10.3)	6.6	16.3	13.5	13.6 ± 7.6 (14.1)	7.1	9.9	8.2	6.8
Stacking	9.4 ± 4.8 (10.0)	5.8	15.8	13.2	13.6 ± 10.4 (13.9)	6.2	9.6	8.0	6.0
Averaging	9.2 ± 4.6 (9.9)	5.1	15.3	12.8	13.3 ± 9.4 (13.7)	5.6	9.5	7.9	5.4
Voting	9.0 ± 4.6 (9.7)	5.0	15.0	12.6	13.1 ± 9.1 (13.5)	5.4	9.3	7.8	5.2

Table 5 Accuracies and mean ranks for all models across each year for Clinic 1
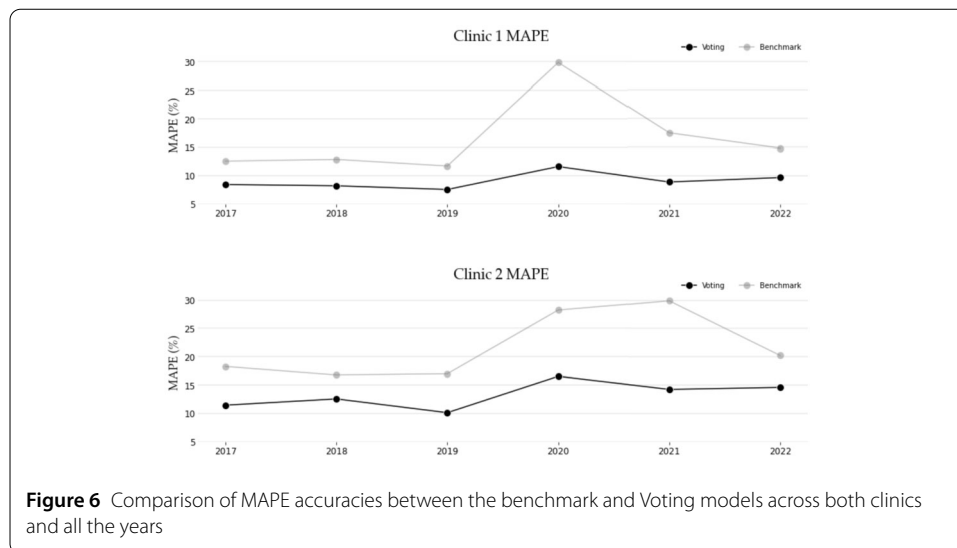
	2017		2018		2019		2020		2021		2022		Mean R
	MAPE	R	MAPE	R	MAPE	R	MAPE	R	MAPE	R	MAPE	R	
kNN	16.8	13.2	16.1	13.3	15.4	13.4	22.7	11.8	16.1	12.2	17.8	12.3	12.7
SVR (NU)	17.2	13.8	15.5	12.9	14.9	13.2	27.2	11.6	15.2	10.9	15.2	11.5	12.4
Benchmark	12.6	10.3	12.9	11.3	11.7	11.8	29.9	12.4	17.5	11.4	14.8	10.8	11.4
Naive	13.4	11.1	12.6	10.9	12.1	11.5	25.6	11.0	18.7	11.1	14.1	9.9	11.0
Naive (Enhanced)	10.0	8.2	9.3	7.5	8.6	7.3	14.9	8.0	11.1	7.7	9.7	6.3	7.6
ARIMA	9.7	7.0	9.7	8.7	9.0	7.7	13.1	7.2	10.8	6.7	10.5	6.8	7.4
Random Forest	9.3	7.7	9.0	7.0	8.1	6.9	13.9	7.3	9.9	7.0	9.8	6.0	7.1
Kernel Ridge	9.0	6.2	9.2	7.1	7.9	5.9	12.5	6.9	10.6	8.2	10.7	8.8	7.0
Ridge	9.0	6.2	9.2	7.2	7.9	6.0	12.4	6.6	10.4	7.7	10.8	8.9	6.9
Prophet	8.4	4.9	8.4	5.8	8.2	7.1	13.5	8.1	10.2	7.3	14.0	9.8	6.9
CatBoost	9.7	8.7	8.7	6.3	8.1	7.0	11.9	5.8	9.7	6.8	9.7	6.0	6.8
Gradient Boosting	9.2	7.5	8.9	6.5	8.0	6.6	12.2	6.1	9.8	6.5	9.7	6.3	6.6
Stacking	8.4	4.9	8.3	5.2	7.8	5.8	12.2	6.5	9.7	6.6	11.1	6.4	5.8
Averaging	8.5	5.5	8.2	4.8	7.6	5.0	11.9	5.4	9.2	5.0	9.9	5.3	5.1
Voting	8.4	4.8	8.2	5.4	7.5	4.8	11.6	5.3	8.9	4.8	9.7	4.8	5.0

Table 6 Accuracies and mean ranks for all models across each year for Clinic 2

	2017		2018		2019		2020		2021		2022		Mean R
	MAPE	R	MAPE	R	MAPE	R	MAPE	R	MAPE	R	MAPE	R	
kNN	17.4	12.2	20.4	13.0	19.5	12.8	27.9	12.2	24.1	11.5	20.2	11.9	12.3
Benchmark	18.3	11.7	16.8	10.3	17.0	11.4	28.2	9.5	29.9	12.7	20.2	10.9	11.1
SVR (NU)	15.4	11.0	19.1	11.8	18.9	12.7	28.4	10.7	22.1	9.8	16.9	9.6	11.0
Naive	16.4	10.4	17.3	10.2	17.9	11.8	23.6	7.6	20.8	11.5	18.9	10.0	10.3
Prophet	12.4	6.5	13.1	6.9	11.2	6.5	23.8	10.0	16.9	8.2	18.3	10.2	7.9
Kernel Ridge	13.2	8.5	13.7	7.9	11.2	7.3	17.4	7.6	14.9	7.0	13.9	6.2	7.5
Ridge	13.0	7.8	13.7	7.8	11.3	7.3	17.5	7.5	14.9	6.7	13.8	6.0	7.3
ARIMA	12.3	6.3	13.8	7.4	11.0	6.2	17.4	8.3	16.4	7.4	16.1	8.0	7.2
CatBoost	12.6	7.7	13.3	7.4	11.2	7.2	14.6	6.2	14.2	6.7	15.6	7.9	7.1
Gradient Boosting	12.7	7.9	13.4	6.8	11.2	7.0	15.7	7.2	14.8	6.8	14.7	6.6	7.1
Random Forest	12.4	7.4	13.4	6.9	11.3	7.4	16.5	6.6	15.2	6.9	15.1	7.2	7.1
Naive (Enhanced)	12.2	6.5	13.3	7.1	11.6	7.1	17.3	7.6	16.4	7.5	13.4	5.3	7.0
Stacking	12.0	6.2	12.6	5.7	10.3	5.2	17.9	6.8	14.5	6.3	15.5	7.6	6.2
Averaging	11.6	5.2	12.6	5.5	10.2	5.1	16.7	6.1	14.2	5.8	15.1	6.8	5.6
Voting	11.5	4.8	12.6	5.3	10.1	5.1	16.5	6.1	14.3	5.4	14.6	5.8	5.4

13.1% accuracies achieved by Clinics 1 and 2 respectively, taking into account that the prior studies did not cover the pandemic period.

To highlight the effects of the pandemic-induced concept drift, Tables 5 and 6 examine the accuracies of the models by each year, across both clinics. Firstly, the tables confirm that the Voting method has consistently generated higher generalisability across all years according to the combined mean rank. However, an abrupt pandemic-induced deterioration in the accuracies can be seen for 2020 onward on both tables. The accuracy declined by 55% and 63% for the Voting method across Clinic 1 and 2 respectively from 2019 to 2020, before beginning to improve from 2021 onward. The subsequent improvement in the accuracies after 2020 indicates that the models made some measure of adjustment to the concept drift and have been able to adapt to the new conditions with the help of the proxy features. In comparison, the benchmark forecasts deteriorated by 156% and 66% across Clinic 1 and 2 respectively for the same period. To visually highlight the effects of the concept drift on the accuracies and the subsequent adaptations by the models, we render the MAPE values for both clinics, contrasting Voting and benchmark models in Fig. 6. Sharp deteriorations especially for the benchmark model can be seen in 2020 for both

Figure 6 Comparison of MAPE accuracies between the benchmark and Voting models across both clinics and all the years

clinics, together with adjustments which occur faster for the Voting model and thus incur a smaller deterioration in accuracy. Though the benchmark models eventually recover some of their accuracy, this occurs with a considerable delay.

Returning to the models' accuracies by year, the overall improvements of the best model (Voting) over the benchmark are ∼33% and ∼23% for Clinics 1 and 2 respectively. A detailed presentation of the percentage improvement achieved by the proposed models over the benchmark approach by year as well as overall can be seen in Tables 8 and 9 in Appendix A. The TheilU statistic comparing the Voting model to the benchmark showed that values are below 1, indicating an improvement over the benchmark. A full summary of these values can be seen in Tables 10 and 11 in Appendix A. Finally, in establishing the advantage of the Voting model over the benchmark, the Diebold-Mariano test across both clinics and for each year indicated that all Voting model forecasts can be considered significantly different (at a 1% level) to the benchmark approach.

Next, we depict an example of the Voting model's forecasting behaviour across three distinctive years, namely from 2019 to 2021, which cover both the stable (2019) and the concept-drift periods (2020–2021) in Fig. 7. This figure shows the forecasting results for estimating patient flows for Clinic 1 while contrasting them with the observed values. Relatively stable patient flows together with accurate forecasts can be seen for 2019, with several undetected patient surges occurring in that year (circled) where the discrepancy between the forecasts and the actuals was 20% or above. Significant COVID-19-related disruptions and the ensuing concept drift can be seen in the figures depicting 2020 and 2021 when recurring lockdowns and GP closures took place. In the figures, the forecasts largely demonstrate a high degree of adaptability to the underlying changes aided by the proposed proxy features which are able to incorporate new information. While the forecasts were generally effective, points of interest have been highlighted in the figures where the observed values exceeded the foretasted values by more than 20%.

To further establish the efficacy of the proxy features, Table 7 contrasts MAPE results of the benchmark and three versions of the Voting model (using only proxy-features, without proxy-features and using a full set of features). The table demonstrates an overall improvement in models using a full set of features that includes proxies features compared to those
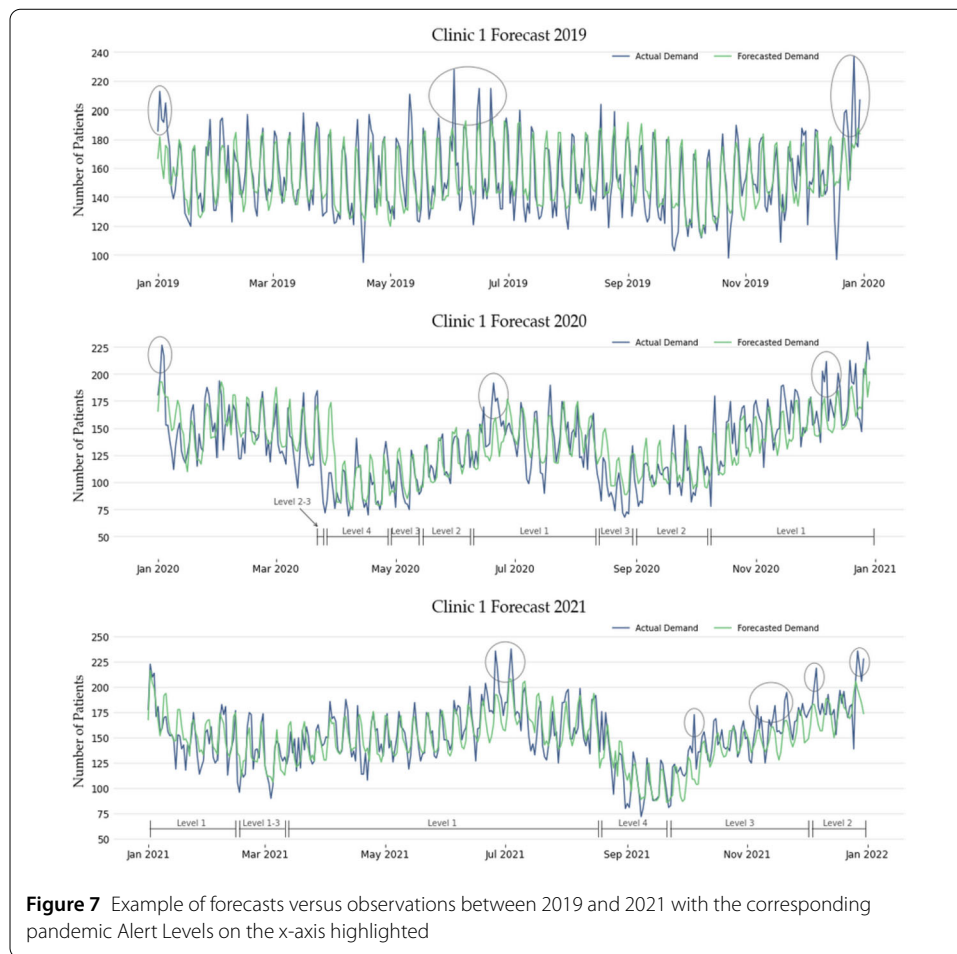
Figure 7 Example of forecasts versus observations between 2019 and 2021 with the corresponding pandemic Alert Levels on the x-axis highlighted

Table 7 Contrasting MAPE accuracies between models using all features versus those generated by proxy-only features for Clinic 1

	2017	2018	2019	2020	2021	2022
Benchmark	12.6	12.9	11.7	29.9	17.5	14.8
Voting (proxy-features only)	11.0	9.7	9.7	12.4	10.0	11.3
Voting (without proxy features)	8.4	8.5	7.7	13.7	10.3	9.2
Voting (all features)	8.4	8.2	7.5	11.6	8.9	9.7

without. The sharp difference between the models is particularly evident for the years 2020 and 2021, which supports the claim that these features are effective in mitigating the effects of concept drift. The results for 2018 and 2019, which are considered normal patterns, also show that models using proxy features perform better than those without. The results for 2017 are inconclusive, and for 2022, the models without proxy features performed better, but it should be noted that the data for 2022 is incomplete, covering only predictions up to July.[8] The contrast between the full-feature model and the one using proxy features only is of note since the latter demonstrates that it was able to account for a considerable amount of variation in the patient flows alone. During stable periods (2017–2019) the full-feature

---

[8]The impact of proxy variables on model forecasts varies throughout the year, with some variables being more effective at different times. Since the flu season onset and severity, as well as other unpredictable factors, vary each year, using only a biased sample of data for the first half of 2022 prevents conclusive findings for the entire year with respect to feature effects.

model improved over the proxy-only model between ∼18% and ∼31%. However, during the concept drift period (2020–2022), this was reduced to only a ∼7% to ∼16% improvement. The experiment indicates that a high degree of information is entailed within the proxy variables and that they are able to explain a significant degree of variation in the dependent variable, especially during the particular case of the recent pandemic-induced concept drift.

4.2 Model interpretability and explainability

We now use the SHAP technique in order to extract the Voting model's interpretability and the explainability of its forecasts for Clinic 1. Figures 8a and 8b depict the high-level interpretability of the model behaviour up to the end of 2019 and 2021 respectively. Again, the two years are selected to highlight the response of the models to the concept drift due to prevailing pandemic conditions affecting patient flows. The figures show the feature importance plots as determined by the models from most to least impactful, while also depicting the relative magnitude of the effect that each feature exerts on the final forecast. The figures also communicate how changes in the values of each feature drive the model's forecast upwards or downwards.

The two figures are in agreement that the most important feature influencing the forecast of patient demand up to the year they represent, are values from seven days prior,
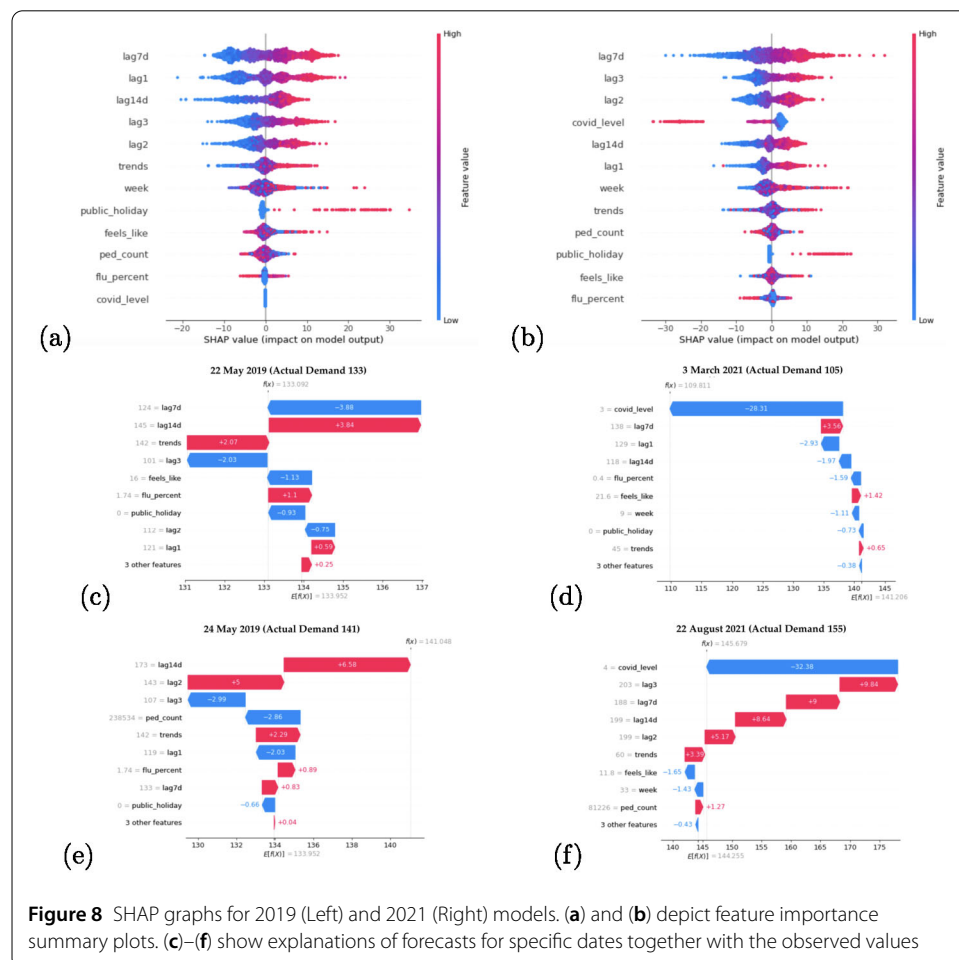
Figure 8 SHAP graphs for 2019 (Left) and 2021 (Right) models. (**a**) and (**b**) depict feature importance summary plots. (**c**)–(**f**) show explanations of forecasts for specific dates together with the observed values

while values from two weeks prior and years are also prominent. However, it is clear that the prevailing COVID-19 Alert Level has gained a high-ranking position for importance in 2021. Generally, the importance of autoregressive features has been interpreted as more important by the models than that of the other proxy variables.

The SHAP summary plots in Figs. 8a and 8b communicate an additional dimension concerning global interpretability. In these figures, one can observe how an increase/decrease in the values of each underlying feature impacts the final forecast. The colours represent feature values where red equates to high values and blue to low values. The x-axis conveys the range of impact on the final forecast. The data points which have a positive SHAP value and appear to the right of the vertical zero line, have an impact on the forecast value which drives it towards predicting higher patient flow numbers. Conversely, the data points that have a negative SHAP value, to the left of the vertical zero line, decrease the forecasted patient flow numbers. As the points extend further from the zero vertical line, their effect and contribution to the final forecast correspondingly increase.

From Figs. 8a and 8b, a general pattern can be seen where the higher values of the features have a correlated positive effect on pushing the forecasts toward higher totals. In Fig. 8b, an exception can be seen in the effects of the COVID-19 Alert Level, where higher levels result in decreasing the forecasted patient flow. The impact of this feature as well as that of the public holidays is also asymmetrical. In other words, the model responds more strongly to forecasting lower patient demand if the values for the COVID-19 Alert Level are high, then it would predict a higher patient flow if the values for these features were low. The reverse holds for the public holiday feature. Some ambiguity exists in the effects that the values of 'feels like', 'pedestrian count' as well as 'flu percent' have on the final prediction.

Figures 8c–f expose the explainability of the models for a selection of specific forecast dates. These figures depict the top nine features for each forecast and their values on the y-axis. The features are rank-ordered by their impact on the final prediction. The figure can be interpreted as a contest of forcing effects between all the features. The expected value of each figure is denoted as $E[f(x)]$ which is the average value of all the forecasted data points. A final SHAP value at the top represents the eventual forecasted outcome. Blue bars represent the forcing effects towards smaller forecast patient flows, while red indicates the opposite. The size of the bars represents the effect size that each feature and its corresponding values exert. These graphs are best interpreted from the bottom up.

The forecasts for Figs. 8c and 8e were made in 2019 which explains why the COVID-19 Alert Level did not play a role in the forecasts. In Fig. 8c, the 7 and 14-day autoregressive features were the most impactful; however, their effects cancelled each other out, while Google search terms pushed the forecasts towards higher estimates. In contrast, Fig. 8e shows that the autoregressive value from two years before had a strong effect on higher forecasts, and pedestrian counts featured more prominently in pushing the final forecast down. Figures 8d and 8f represent the concept-drift phase. In these figures the strong influence of the COVID-19 Alert Level can be seen responding to the changes in the dependent variable, pushing both forecasts down due to the high prevailing pandemic situation. In both cases, the pandemic level eclipsed the remaining features and successfully rectified the final forecast toward an accurate estimation.

Here, we revisit the proxy-only models to study their effects and utility, which we conduct with the removal of the autoregressive features. Figures 9a–f depict the behaviour of
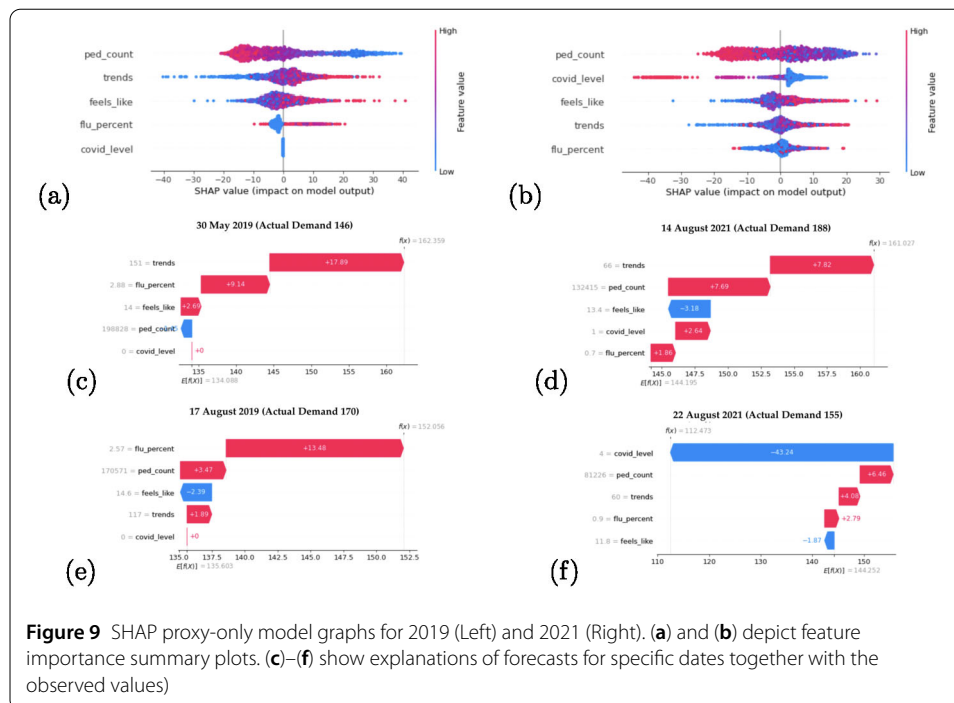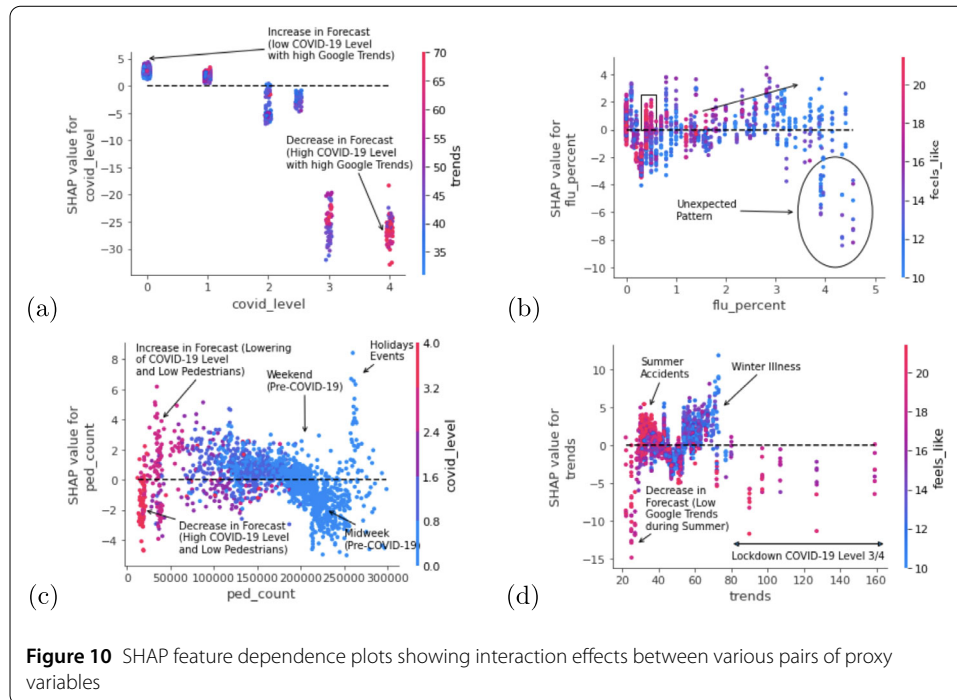
Figure 9 SHAP proxy-only model graphs for 2019 (Left) and 2021 (Right). (**a**) and (**b**) depict feature importance summary plots. (**c**)–(**f**) show explanations of forecasts for specific dates together with the observed values)

these models. The same analysis approach is followed as previously by using data up to 2019 and 2021. Again, the first two Figs. 9a and 9b depict the interpretability of the models. In both years, it can be seen that pedestrian foot traffic is a significant driver of the forecasts being negatively correlated with the forecasted patient flows. Google trends and 'feels like' data similarly exert a positively correlated influence across both years; however, for 2021 the COVID-19 Alert Level becomes the second most important feature with a negatively correlated impact on the patient flows.

Figures 9c and 9e depict the explainability of the forecasts for two separate dates in 2019. Both explanatory figures demonstrate the strong influence of the prevailing influenza conditions as a strong variable for this period, while Google trends searches are the strongest for the former. For 2021, we observe that Google trends search results, as well as pedestrian foot traffic, are influential on both dates in Figs. 9d and 9f; however, the COVID-19 Alert Level is an overwhelmingly influential variable for the latter in driving the forecasts down due to the highest prevailing pandemic level of 4 occurring in Fig. 9f. The different magnitudes of effect that the COVID-19 Alert Level is explained by Fig. 9b, where we see that high alert levels exert a considerably larger effect on the forecast than lower alert levels exert on increasing the forecasts.

We conclude with a high-level analysis of the model's behaviour to draw out further insights, but this time considering how pairs of features interact in order to influence the final forecasts. We depict a selection of these interactions in Figs. 10a to 10d focusing primarily on proxy features. Each plot contains a horizontal line which indicates the threshold representing a transition from increasing and decreasing effects that pairs of features have on the final forecast

Figure 10a shows how the effect on the final forecast changes as the COVID-19 Alert Level increases (x-axis) and as the value for Google Trends search terms rises (colour gradient on the right y-axis). The interaction of low COVID-19 Alert Level values and high

Figure 10 SHAP feature dependence plots showing interaction effects between various pairs of proxy variables

Google Trends search terms exerts the strongest effect on raising the forecasts. The forecasts are significantly downgraded as the COVID-19 Alert Level values increase and the lockdown measures take effect but at the same time, some of the highest Google Trends search terms regarding respiratory terms are encountered during the highest lockdown conditions.[9]

Several insights can be extracted from Fig. 10b depicting the interaction between the reported increase in influenza prevalence and the weather variable. Firstly, it can be observed that influenza levels between 0.5% and 1% in conjunction with high weather values tend to correspond with positive effects on the forecasts. This can be explained by summer conditions where there is an uptick in sport-related injuries and general outdoor accident-based cases. However, influenza levels below 0.5% tend to be inconsistent. As the influenza prevalence increases from 1% to 3%, this has a positive effect on forecasts as would be expected during the winter months, and can be confirmed by the fact that most of the weather-perception values are downward trending. An unexpected pattern can be seen as influenza prevalence levels increase beyond 3% where counter-intuitively a mixed, and more negative effect is exerted on the forecasts which is contrary to the reported experience of the clinic's staff.[10]

Meanwhile, in Fig. 10c it is observable that as the pedestrian traffic increases, there is a generally a decreasing effect on the forecasts[11] until the highest pedestrian traffic values

---

[9]The greatest values indeed coincided with the first and the highest COVID-19 Alert Level announced in March 2020.

[10]The negative effect of high influenza prevalence may be due to uncertainty in the data itself which raises some questions about the reliability of this feature since it is incomplete with the data being available from 2018 onward. It may also be explained by the fact that this feature is a nationwide indicator and influenza incidence is likely to be different in colder regions of the country than in Auckland. The fact that the highest values of flu prevalence coincide with high weather values as negative drivers of forecasts may be suggestive that this period corresponds with the earlier beginning of spring in Auckland than in other regions, which leads to a sharp actual decline in respiratory related illness.

[11]This is explainable by the fact that the pedestrian traffic peaks midweek when the patient arrivals are at their lowest. Meanwhile, the pedestrian traffic is lowest in the CBD during the weekend when patient demand is typically at its highest.

are reached, at which point, the behaviour of the models becomes more erratic, possibly due to public holidays and large events. The highest COVID-19 Alert Level values together with the ensuing lowest pedestrian traffic values, aggressively decrease the forecasts as expected. The highest positive effects on the forecasts are seen in data points where the pedestrian traffic is low but the COVID-19 Alert Level is less than 4 which corresponds with the most stringent lockdown mandates. This also corresponds with the loosening of the lockdown restrictions which in turn trigger increased clinic visits due to the pent-up demand during the highest lockdown conditions.

In the last illustration, Fig. 10d depicts the effects of increasing Google Trends search terms with respect to changes to weather conditions. We see that some of the highest Google Trends search values (from 80 onward) correspond with the strongest negative forecasting effects. We can correlate this result with Fig. 10 which shows that some of the highest Google Trends values occurred during the highest lockdown measures. A strong negative effect on the forecasts can be explained by the increase in the general anxiety levels of the homebound populace, resulting in an increase in Google searches covering terms related to respiratory illnesses, while at the same time discouraging the population from seeking non-urgent medical attention to lessen the risk of exposure as directed by the high covid alert levels. Therefore, it can be surmised that unusually high Google Trends search values from 80 onward are exceptional as they are attributable to severe lockdown conditions. Meanwhile, Google Trends search values ranging from 50 to 80 can be interpreted as belonging to the winter months and represent typical patterns that are confirmed by the low weather temperatures for this range. The effect of the summer months and the lowest Google Trends search values can also be observed where the strongest negative effects on the forecast values are exerted.

5 Discussion

A comprehensive empirical investigation was conducted using a broad set of models and features which achieved considerably improved accuracies over the benchmark approaches as well as competitive accuracies with respect to prior studies. In addressing the study's first research question (RQ1), we find that the ensemble-based Voting method consistently outperformed all other candidate algorithms for this setting. The approach successfully combined a mixture of both machine learning and statistical techniques in order to leverage the advantages of both. This finding aligns with the literature on ensemble-based machine learning which attests to the benefits of using such methods in comparison to utilizing single models. The theory of ensemble-based machine learning posits that by aggregating the predictions of diverse models, the resulting ensemble will on average exhibit superior performance across a wide range of datasets when compared to utilizing a single model.

The abrupt onset of the concept drift caused by the COVID-19 pandemic conditions degraded the forecasting accuracies initially. Nonetheless, the experiments indicated that the suite of novel real-time proxy variables were effective at adjusting the models to the new and evolving drivers of patient flows. Not only were the proposed features useful at triggering adaptations to the disruptions of historic patterns, but they were also effective at improving the model forecasts in general (RQ2). Indeed, the experiments showed that reasonable models could be generated using only proxy variables without reliance on autoregressive feature values. The study confirmed that the COVID-19 Alert Level feature

was particularly effective at forecasting patient volumes during disruptive periods, while determining the effectiveness of the Google search terms, pedestrian foot traffic as well as weather data. While the reported prevalence of influenza prevalence was a useful feature and contributed to some degree in improving forecasts, it was less reliable than the other proxy features likely due to its incompleteness. The implications of this finding is that accurate forecasting of patient flows should remain possible using these features in the event of future pandemic outbreaks.

A particularly novel contribution of this study was the extensive use of XAI tools in order to expose the internals of the forecasting models. The use of the SHAP technique to achieve global interpretability of the models was highlighted and proved effective in depicting the various effects that features and their values bear on the final forecasts (RQ3). Additionally, at the local level, the technique demonstrated its ability to clearly explain the drivers of the forecasted values for individual days. Meanwhile, the dependence plots yielded extra insights into the interactions of various features and how they collectively influence the final forecasts. This also enabled the validation of the models against domain knowledge to take place.

The findings of this study contain practical implications for practitioners. While the results of this study are specific to the two clinics in question, replication of our approach to other clinics in different cities would require re-evaluation of the features used. However, the insights and findings of this study can inform practitioners about the potential utility of these features for generalisation and their potential responsiveness to concept drifts. Additionally, the study's results can inform practitioners about the potential performance and limitations of different machine learning algorithms when applied to patient flow prediction tasks. Our analysis approach is therefore generalisable to different contexts, and the demonstrated XAI tools can be utilised to evaluate and fine-tune models. We offer some recommendations to practitioners:

- Consider the timeliness and granularity of data: Choose real-time proxy variables that are available in a timely manner and at the appropriate level of granularity for accurate forecasting, ensuring that there also exists a sufficient amount of historic values for model training.
- Consider unique attributes of the clinic: When selecting real-time proxy variables for patient forecasting, consider the unique features of the clinic, such as its location, the local transportation system and nearby hospitals.
- Explore diverse data sources: Look beyond traditional data sources, and explore other sources of data, such as social media, Google Analytics, ED capacities at nearby hospitals and local General Practitioners if available.
- Consider using ensemble methods: Model forecasts using ensemble methods, such as combining multiple algorithms or models, to improve forecasting accuracy.
- Evaluate and detect concept drift: Simple concept drift methods that track the error rate of a model can be applied for this purpose where drift is detected when the error rate increases significantly compared to a minimum observed error rate. Confidence intervals can be used to determine if a significant increase in error rate has occurred.
- Anticipate and mitigate for concept drift: Stay ahead of some causes of concept drift which can come in the form of changes in nearby hospital operations of their EDs, competition from new UCCs and reforms in healthcare policies concerning subsidies. Concept drifts can also accompany infrastructure and transportation changes as well

as economic downturns. While these factors can lead to fluctuations in patient arrival patterns and demand for services, exploratory modelling using additional flag variables that attempt to represent these events may act to preserve forecasting accuracy.

6 Conclusion

Patient flows in Urgent Care Clinics (UCCs) and Emergency Departments (EDs) have been experiencing increasing pressure as the volumes have grown and have become less predictable using *ad hoc* approaches. The ability to accurately forecast patient arrivals in these contexts, and to understand better what some of the drivers of demand are, is important for the efficient and effective functioning of these healthcare providers by enabling them to respond faster and achieve more optimal human resourcing.

There are many factors which affect patient arrivals and randomness accounts for some of this. Autoregressive variables as well as calendar and meteorological indicators have already been explored in prior research and have been found useful for estimating patient flows. This study goes further by also considering additional quasi-real-time variables like Google search terms, pedestrian traffic and prevailing incidence levels of influenza, alongside COVID-19 Alert Level indicators to improve the forecast accuracy.

This study makes a unique contribution in several respects. Not only does this study integrate a wide variety of new features for forecasting patient flows, but it also considers both the effects that the concept drift triggered by the recent pandemic conditions had on the forecast accuracies, and the means by which the models can rapidly adapt to the changing context in order to learn and recover. This research also ventured beyond the standard studies using machine learning methods in this domain, by utilising tools from the eXplainable AI field in order to expose the internals of the models, thus achieving both the high-level interpretability of the models and the explainability of their individual forecasts.

Our research determined that the Voting ensemble-based method performed most reliably in this setting and the final accuracies were competitive with those from prior studies. While the autoregressive and calendar features were important, the experiments indicated that the prevailing COVID-19 Alert Level feature together Google search terms and pedestrian traffic were effective at generating accurate forecasts.

Appendix A: Detailed model accuracies and statistical test results
A.1 Quantifying improvements of proposed methods versus the benchmark

Tables 8 and 9 show the improvements achieved by the top seven proposed methods using all features over the benchmark as a percentage of MAPE. The results are broken down by year.

A.2 TheilU statistical test on the significance of model improvements over the benchmark

Tables 10 and 11 display the results of the TheilU statistical test showing the detailed improvements attained by the top seven proposed methods using all features over the benchmark. The test is performed across all years as a summary as well as broken down by year

Table 8 Percentage improvement of proposed models over the benchmark for Clinic 1 across all years

	2017	2018	2019	2020	2021	2022	Overall
Prophet	26.8	29.8	25.5	27.7	26.1	3.3	25.0
Random Forest	17.1	26.2	27.8	38.5	29.7	22.2	27.3
CatBoost	13.6	27.6	26.5	43.5	30.5	25.3	28.1
Gradient Boosting	18.4	26.9	28.5	43.1	31.3	24.6	29.2
Stacking	25.7	30.6	30.1	40.5	31.0	19.1	30.5
Averaging	24.6	31.3	32.1	42.0	34.5	25.4	32.2
Voting	25.5	31.0	32.3	41.2	36.1	27.2	32.7

Table 9 Percentage improvement of proposed models over the benchmark for Clinic 2 across all years

	2017	2018	2019	2020	2021	2022	Overall
Prophet	25.8	12.0	26.2	−25.0	29.6	−15.1	11.2
Gradient Boosting	21.6	12.2	28.7	−0.7	38.0	17.2	19.7
Random Forest	23.8	13.0	27.3	2.0	36.1	19.1	20.3
CatBoost	23.0	12.0	26.9	5.1	38.1	14.0	20.4
Stacking	27.5	17.0	33.7	−2.6	37.9	13.6	21.9
Averaging	29.0	16.8	34.0	1.1	39.4	16.1	23.4
Voting	29.8	16.6	34.6	−1.2	40.3	17.2	23.4

Table 10 TheilU Statistic for Clinic 1

	2017	2018	2019	2020	2021	2022	Overall
Random Forest	0.576	0.623	0.635	0.99	0.894	0.989	0.762
Prophet	0.507	0.584	0.612	0.927	0.873	1.417	0.756
Gradient Boosting	0.566	0.622	0.63	0.863	0.872	0.965	0.721
CatBoost	0.598	0.599	0.629	0.85	0.866	0.959	0.719
Stacking	0.516	0.573	0.596	0.845	0.865	1.127	0.706
Averaging	0.525	0.565	0.588	0.833	0.816	1.000	0.685
Voting	0.516	0.559	0.581	0.821	0.795	0.956	0.671

Table 11 TheilU Statistic for Clinic 2

|	2017	2018	2019	2020	2021	2022	Overall
Prophet	0.774	0.76	0.591	1.584	1.064	1.133	1.045
Stacking	0.725	0.693	0.56	1.167	0.888	0.982	0.855
Random Forest	0.743	0.73	0.631	1.026	0.962	0.969	0.845
Averaging	0.703	0.689	0.566	1.052	0.871	0.967	0.815
Gradient Boosting	0.736	0.732	0.626	0.894	0.916	1.043	0.810
Voting	0.692	0.684	0.558	1.009	0.886	0.947	0.800
CatBoost	0.754	0.737	0.63	0.836	0.872	1.069	0.796

for each clinic. The tables show that all the values for the Voting model are below 1, indicating an improvement over the in-house method, apart from one instance in 2020 for Clinic 2, where the performance was comparable.

A.3 Accuracies of the proxy-only model

Table 12 shows the detailed MAPE accuracies of the proxy-only models for Clinic 1 contrasted with univariate ARIMA and Prophet models. The univariate models outperform the proxy-only models in the period leading up to the pandemic, but subsequently the Voting proxy-only models shows general improvements over the univariate methods as the concept drift takes effect from 2020 onward.

Table 12 MAPE accuracies for the proxy-only models for Clinic 1

	2017		2018		2019		2020		2021		2022		Overall mean	
	MAPE	R	MAPE	R	MAPE	R	MAPE	R	MAPE	R	MAPE	R	MAPE	R
CatBoost	16.9	4.5	14.0	4.1	14.0	4.1	16.5	3.7	12.8	3.6	14.0	3.8	14.8	4.0
Random Forest	16.3	4.2	14.0	4.2	14.3	4.3	16.3	3.7	13.0	3.8	13.3	3.3	14.6	4.0
ARIMA	9.7	2.0	9.7	2.5	9.0	2.3	13.1	2.6	10.8	2.7	10.5	2.4	10.5	2.4
Prophet	8.5	1.7	8.4	1.9	8.4	2.1	13.3	2.8	10.2	2.6	14.2	3.5	10.2	2.3
Voting	11.0	2.6	9.7	2.3	9.7	2.3	12.4	2.1	10.0	2.2	11.3	2.0	10.6	2.3
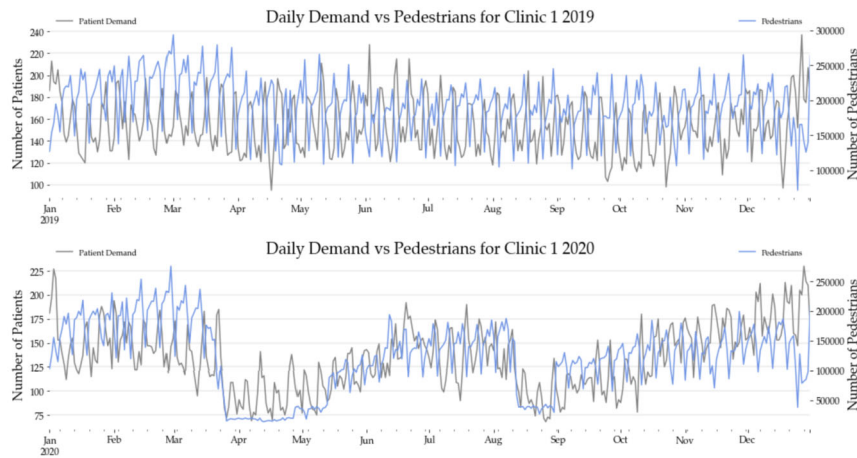
Figure 11 Pedestrian Traffic Proxy Feature for 2019 and 2020
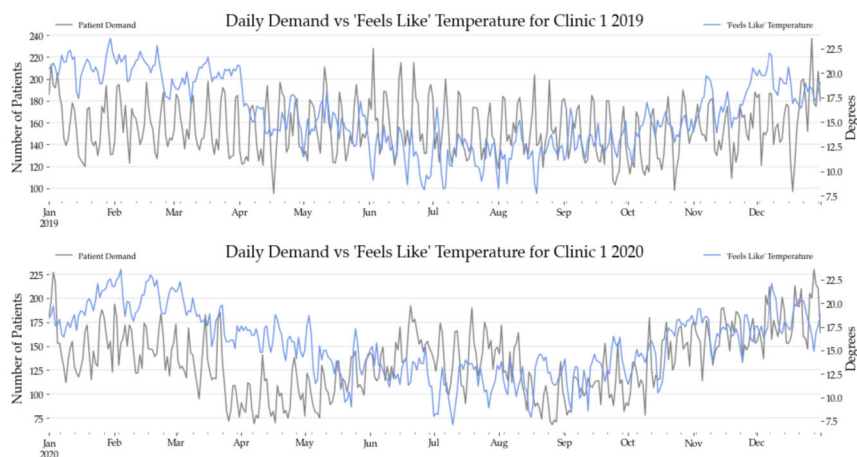
Figure 12 'Feels Like' Temperature Proxy Feature for 2019 and 2020

Appendix B: Proxy feature values versus patient flows

Figures 11 to 14 highlight the correlation between the patient arrivals per day for Clinic 1 against the proxy variable values for 2019 and 2020, contrasting trends over stable and concept drift periods.
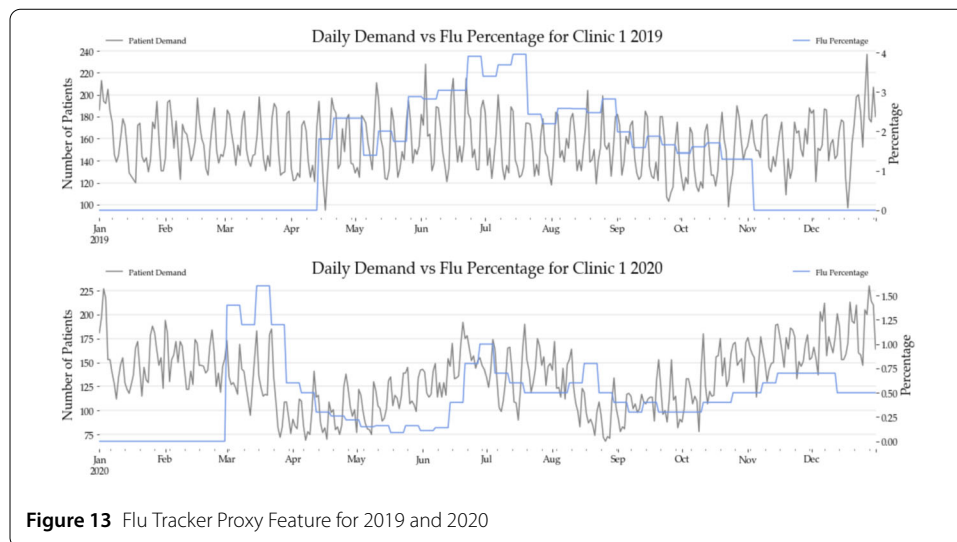
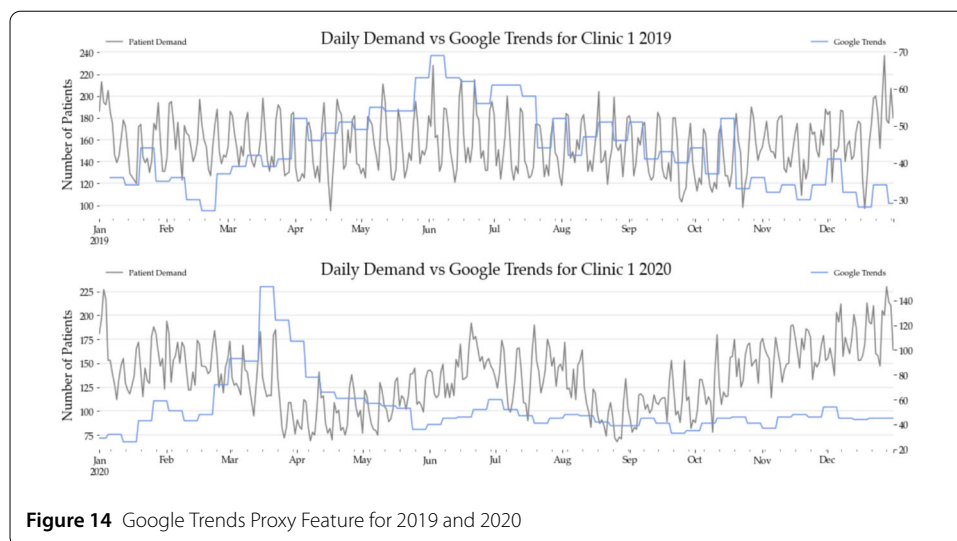**Figure 13** Flu Tracker Proxy Feature for 2019 and 2020

Figure 14 Google Trends Proxy Feature for 2019 and 2020

Availability of data and materials
The data for this study cannot be made available due to its sensitivity.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
The authors contributed equally to this research. PM performed data acquisition, cleaning and the writing of all software for conducting the experiments and the implementation of the models. TS conceptualised the aims and the direction of the study and wrote the initial draft. Both authors analysed the results as well as contributed to the final manuscript, and approved its submission for publication.

References

1. Aboagye-Sarfo P, Mai Q, Sanfilippo FM, Preen DB, Stewart LM, Fatovich DM (2015) A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in western Australia. J Biomed Inform 57:62–73.
2. Batal H, Tench J, McMillan S, Adams J, Mehler PS (2001) Predicting patient visits to an urgent care clinic using calendar variables. Acad Emerg Med 8(1):48–53.
3. Bowerman BL, O'Connell RT, Koehler AB (2005) Forecasting, time series, and regression: an applied approach, vol 4. South-Western Pub.
4. Box GE, Jenkins GM (1976) Time series analysis: forecasting and control. Holden-Day, San Francisco
5. Boyle J, Jessup M, Crilly J, Green D, Lind J, Wallis M, Miller P, Fitzgerald G (2012) Predicting emergency department admissions. Emerg Med J 29(5):358–365
6. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
7. Calegari R, Fogliatto FS, Lucini FR, Neyeloff J, Kuchenbecker RS, Schaan BD (2016) Forecasting daily volume and acuity of patients in the emergency department. Comput Math Methods Med 2016:3863268
8. Carvalho-Silva M, Teresa M, Monteiro T, de Sá-Soares F, Dória-Nóbrega S (2018) Assessment of forecasting models for patients arrival at emergency department. Oper Res Health Care 18:112–118
9. Champion R, Kinsman LD, Lee GA, Masman KA, May EA, Mills TM, Taylor MD, Thomas PR, Williams RJ (2007) Forecasting emergency department presentations. Aust Health Rev 31(1):83–90.
10. Coster JE, Turner JK, Bradbury D, Cantrell A (2017) Why do people choose emergency and urgent care services? A rapid review utilizing a systematic literature search and narrative synthesis. Acad Emerg Med 24(9):1137–1149
11. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27
12. Cristianini N, Shawe-Taylor J et al (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
13. DeLurgio S, Denton B, Cabanela RL, Bruggeman S, Williams AR, Ward S, Groves N, Osborn J (2009) Forecasting weekly outpatient demands at clinics within a large medical center. Prod Invent Manag J 45(2):35–46
14. Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13(3):253–263
15. Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V (1996) Support vector regression machines. In: Advances in neural information processing systems, vol 9
16. Fakhfakh K, Ben-Othman S, Jourdan L, Smith G, Renard J-M, Hammadi S, Zgaya H (2021) Ontology for overcrowding management in emergency department. In: MedInfo medical and health informatics
17. FluTracking (2022) Flutracking reports (New Zealand). https://info.flutracking.net/reports-2/new-zealand-reports/
18. Google Trends (2022) Explore what the world is searching. https://trends.google.com
19. Gramegna A, Giudici P (2021) Shap and lime: an evaluation of discriminative power in credit risk. Front Artif Intell 4:752558
20. Guyeux C, Bahi JM (2022) How to predict patient arrival in the emergency room. In: World conference on information systems and technologies. Springer, Berlin, pp 600–610
21. Harrou F, Dairi A, Kadri F, Sun Y (2022) Effective forecasting of key features in hospital emergency department: hybrid deep learning-driven methods. Mach Learn Appl 7:100200
22. He J, Hou X, Toloo S, Patrick JR, Gerald GF (2011) Demand for hospital emergency departments: a conceptual understanding. World J Emerg Med 2(4):253
23. Heart of the City Auckland (2022) We're Auckland's city centre business association: we've got a heart for business. https://www.hotcity.co.nz/
24. Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67
25. Hollander JE, Sharma R (2021) The availablists: emergency care without the emergency department. NEJM Catalyst Innov Care Deliv 2(6)
26. Jones SS, Thomas A, Evans RS, Welch SJ, Haug PJ, Snow GL (2008) Forecasting daily patient volumes in the emergency department. Acad Emerg Med 15(2):159–170.
27. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems, NIPS'17. Curran Associates, Red Hook, pp 4768–4777
28. Maddigan P, Susnjak T (2022) Forecasting patient demand at urgent care clinics using machine learning. arXiv preprint. arXiv:2205.13067
29. Marcilio I, Hajat S, Gouveia N (2013) Forecasting daily emergency department visits using calendar variables and ambient temperature readings. Acad Emerg Med 20(8):769–777
30. Mathrani A, Susnjak T, Ramaswami G, Barczak A (2021) Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. Comput Educ Open 2:100060
31. Minh D, Wang HX, Li YF, Nguyen TN (2022) Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 55:3503–3568
32. Navares R, Díaz J, Linares C, Aznarte JL (2018) Comparing arima and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid. Stoch Environ Res Risk Assess 32(10):2849–2859
33. New Zealand Government (2022) History of the COVID-19 alert system. https://covid19.govt.nz/about-our-covid-19-response/history-of-the-covid-19-alert-system/
34. New Zealand Government (2022) Traffic lights. https://covid19.govt.nz/traffic-lights/
35. Pearson K (1905) The problem of the random walk. Nature 72(1865):294
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E, (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
37. Petsis S, Karamanou A, Kalampokis E, Tarabanis K (2022) Forecasting and explaining emergency department visits in a public hospital. J Intell Inf Syst 59:479–500
38. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2017) Catboost: unbiased boosting with categorical features

39. Rocha CN, Rodrigues F (2021) Forecasting emergency department admissions. Intell Data Anal 25(6):1579–1601
40. Sahu SK, Baffour B, Harper PR, Minty JH, Sarran C (2014) A hierarchical Bayesian model for improving short-term forecasting of hospital demand by including meteorological information. J R Stat Soc, Ser A, Stat Soc 177(1):39–61
41. Shapley LS (1953) Quota solutions op n-person games. Edited by Emil Artin and Marston Morse p 343
42. Sudarshan VK, Brabrand M, Range TM, Wiil UK (2021) Performance evaluation of emergency department patient arrivals forecasting models by including meteorological and calendar information: a comparative study. Comput Biol Med 135:104541
43. Susnjak T, Barczak ALC, Hawick KA (2012) Adaptive cascade of boosted ensembles for face detection in concept drift. Neural Comput Appl 21(4):671–682
44. Taylor SJ, Letham B (2018) Forecasting at scale. Am Stat 72(1):37–45
45. Theil H (1961) Economic forecasts and policy
46. Visual Crossings (2022) Weather data & API: global forecast & history data. https://www.visualcrossing.com/
47. Vollmer MAC, Glampson B, Mellan T, Mishra S, Mercuri L, Costello C, Klaber R, Cooke G, Flaxman S, Bhatt S (2021) A unified machine learning approach to time series forecasting applied to demand at emergency departments. BMC Emerg Med 21(1):9.
48. Wargon M, Guidet B, Hoang TD, Hejblum G (2009) A systematic review of models for forecasting the number of emergency department visits. Emerg Med J 26(6):395–399. https://emj.bmj.com/content/26/6/395.
49. Whitt W, Zhang X (2019) Forecasting arrivals and occupancy levels in an emergency department. Oper Res Health Care 21:1–18
50. Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259
51. Xu M, Wong T-C, Chin K-S (2013) Modeling daily patient arrivals at emergency department and quantifying the relative importance of contributing variables using artificial neural network. Decis Support Syst 54(3):1488–1498
52. Xu Q, Tsui K-L, Jiang W, Guo H (2016) A hybrid approach for forecasting patient visits in emergency department. Qual Reliab Eng Int 32(8):2751–2759
53. Zhang Y, Zhang J, Tao M, Shu J, Zhu D (2022) Forecasting patient arrivals at emergency department using calendar and meteorological information. Appl Intell 52:11232–11243

Publisher's Note