



A Temporal Fusion Transformer for Long-Term Explainable Prediction of Emergency Department Overcrowding

Francisco M. Caldas^(✉)  and Cláudia Soares 

NOVA School of Science and Technology, Caparica, Portugal
f.caldas@campus.fct.unl.pt, claudia.soares@fct.unl.pt

Abstract. Emergency Departments (EDs) are a fundamental element of the Portuguese National Health Service, serving as an entry point for users with diverse and very serious medical problems. Due to the inherent characteristics of the ED, forecasting the number of patients using the services is particularly challenging. And a mismatch between the affluence and the number of medical professionals can lead to a decrease in the quality of the services provided and create problems that have repercussions for the entire hospital, with the requisition of health care workers from other departments and the postponement of surgeries. ED overcrowding is driven, in part, by non-urgent patients, that resort to emergency services despite not having a medical emergency and which represent almost half of the total number of daily patients. This paper describes a novel deep learning architecture, the Temporal Fusion Transformer, that uses calendar and time-series covariates to forecast prediction intervals and point predictions for a 4 week period. We have concluded that patient volume can be forecasted with a Mean Absolute Percentage Error (MAPE) of 5.91% for Portugal's Health Regional Areas (HRA) and a Root Mean Squared Error (RMSE) of 84.4102 people/day. The paper shows empirical evidence supporting the use of a multivariate approach with static and time-series covariates while surpassing other models commonly found in the literature.

Keywords: Time series · Emergency department · Machine learning · Temporal fusion transformer · Forecasting · Manchester triage system · Neural network · Explainable ML · National Health Service

1 Introduction

The forecast of the number of patients who use emergency services daily is essential to determine in advance the human resources needed at hospital Emergency Departments (ED). Multi-step ahead predictions allow hospital managers

This work was partially supported by the strategic project NOVA LINC'S (UIDB/04516/2020), the FCT project DSAIPA/AI/0087/2018 and the Carnegie Mellon University - Portugal FCT project CMU/TIC/0016/2021.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
I. Koprinska et al. (Eds.): ECML PKDD 2022 Workshops, CCIS 1752, pp. 71–88, 2023.
https://doi.org/10.1007/978-3-031-23618-1_5

to organise rotation schedules and diminish waiting times in urgent care facilities [21,39]. When not accounted for, overcrowding can lead to a decrease in the quality of patient care and worse clinical outcomes [5,20]. From a macro point of view, the influx in the emergency department combines an expected number of people who are taken to the emergency room with a very serious illness, for example, heart attack, with people that use the emergency hospital to deal with non urgent problems, such as common cold, strained muscles, or to deal with problems associated with chronic illness [20,42]. The most serious cases are reasonably constant over time, and, predominantly, people in life threatening conditions have no choice but to go to emergency care, thus the indicators of a rise in patients with serious illnesses might not be the same for non urgent users. A large number of patients that resort to urgent care are not, however, urgent, according to the Manchester Triage system, used in the Portuguese National Healthcare System. Roughly 40% of the patients are classified during triage at the green/blue level, which means not urgent. Unlike more urgent patients, the influx of green/blue patients has several factors that follow well-defined cycles. For example, it is easy to identify that the day with the most influx of non-urgent patients is Monday, with a smaller number of patients pursuing emergency care during the weekend [4,19,32]. To combine the predictive power of Deep Neural Networks with the explainability usually reserved for simpler algorithms, we will use a recently developed machine learning model to predict the influx of non-urgent patients: the Temporal Fusion Transformer (TFT)[26]; and study which variables, time-series or not, had the most impact on the model, and thus which are most relevant to predict daily patient volume.

In the following section, we will perform a brief literature review of the work done to tackle this problem, followed by a section in which we display the data and offer some exploratory analysis to obtain a better understanding of the dataset. In Sect. 4, the methodology of the experiment will be displayed, presenting the goals, the forecast horizon, and the forecast model. In the results section, we perform a comparison of the TFT model with other known models in the literature, followed by an analysis of covariate importance and attention weights. Finally, the conclusions of this study are drawn, acknowledging the strengths and limitations of the TFT model, and proposing future work.

2 Literature Review

Previous studies have examined the multi-step forecasting of daily patient volumes [10,21]. Most focus is on the use of classical statistical tools for temporal linear regression such as moving averages [28], and their many extensions, namely ARIMA, SARIMA or VARIMA [3,7,34,40]. In recent years, with the advent of machine learning, newer studies have been conducted that use neural networks [21,43], or otherwise other machine learning techniques to tackle the same problem [29,33,37]. From the use of Feed-forward Neural Networks [21,29], to Recurrent Neural Networks [15,22], 1-D Convolution Neural Networks [35], and later to Long Short-Term Unit (LSTM) [15,36], there has been a constant advance

in the field, from linear models to deep neural network models. In most studies using ARIMA and its variants, it was found that calendar variables (day, day of the week, holidays) have a significant contribution to model results [6, 17, 21, 39]. Weather data, such as temperature and rain, have shown predictive power for ED arrivals with respiratory problems [29], but in others studies that analysed the whole spectrum of ED visitors, it is either not a significant variable, or it could be replaced by calendar variables, e.g. month of the year [17, 39]. This level of covariate interpretability is one of the frequent drawbacks of Neural Networks, alongside the failure to recognize long-term dependencies in time-series. One specific device that addresses both problems is the Attention Mechanism [38]: simply put, it evaluates long-term dependencies and also represents how each time-step impacts the model’s prediction. Attention has been used as part of a specific Neural Network family of architectures called Transformers, that has shown impressive results in the Natural Language Processing field [9, 41]. In the literature, we found only one example that used a Temporal Fusion Transformer model to predict Emergency Department (ED) volume in one hospital for one day ahead [31]. While not being the only work that performed only daily predictions [33, 37], we find that a longer forecasting window produces increased value for hospital management and poses a different challenge from a machine learning perspective, as seasonal fluctuation needs to be fully represented, and common forecasting models tend to decrease in predictive quality as the forecast period becomes wider.

3 Data Analysis

In this section, we will present the database used in this work. The data was obtained from the public database “Transparência SNS”¹ and refers to daily data of care in primary health centres together with daily data of consultations and waiting times in hospitals’ emergency departments (ED) across Portugal, divided by Regional Health Area (RHA). The time analyzed covers the time period from November 1st, 2016 to February 20th, 2022; with 6353 individual observations and 16 variables per observation that define, among other things, the Regional Health Area (RHA), *Área Regional de Saúde (ARS)* in Portuguese, of the observation. In total, the dataset contains information regarding the daily volume of patients in emergency care, the number of scheduled and unscheduled consultations in primary health facilities, the daily number of patients arriving at the Emergency Department (ED) with respiratory issues, the waiting times between triage and the first medical evaluation and categorical variables pertaining to calendar information, such as weekend, day of week or national holidays.

In Fig. 1 we can observe the weekly variation in the number of non-urgent patients, as well as the volume shift according to RHA. It is visible that despite having different levels of affluence, the different RHA follow the same trend, with peaks of affluence occurring on Monday, and reduced volume on weekends

¹ <https://transparencia.sns.gov.pt/explore/dataset/atividade-sindrome-gripal-csh>
<https://transparencia.sns.gov.pt/explore/dataset/atendimentos-nos-csp-gripe>.

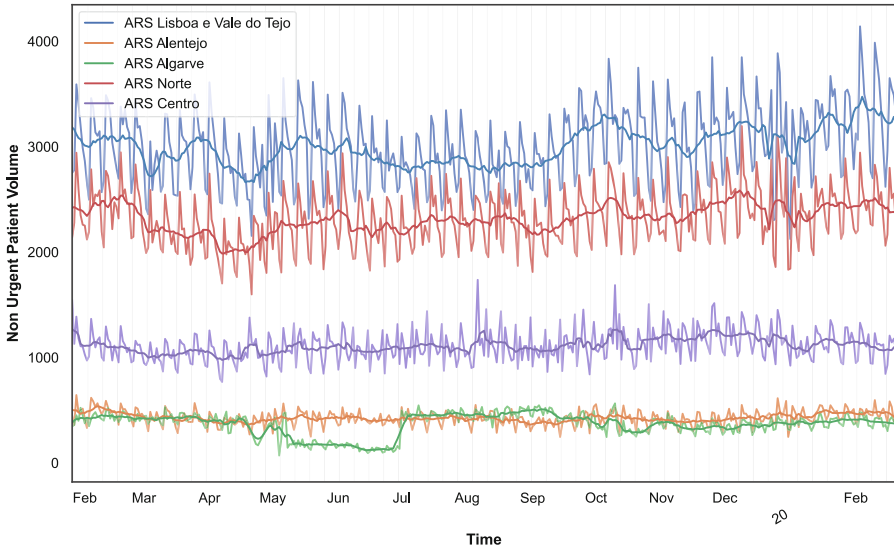


Fig. 1. Time series from January, 2019 to February 20, 2020. We can observe the weekly cycle, as well as annual trends and volume variation according to Regional Health Area. The weekends are marked with grey lines, corresponding with diminishing number of non urgent patients searching for emergency care.

and during the Summer months, usually associated with vacations. This aspect of the data served as motivation for the application of a non-linear model over multiple time series, unlike well-established models such as ARIMA.

Another interesting feature of the data is the observation of the period in which Portugal was affected by the COVID-19 disease and took containment measures that reduced travel and in person work: in this period (10/03/2020–1/08/2021) the percentage of non urgent visits in the RHA of *Lisboa e Vale do Tejo* dropped from the normal value of 48% of the total to 40%, with more dramatic drops for example in the *Algarve* RHA from 45% to 30% at the beginning of the pandemic. This dramatic period influenced the way people used emergency services, and it can demonstrate how external factors influence people going to the emergency room. This shift, associated with the general decline in the number of people in urgent care, urgent or not, represents a distribution change in the time series, therefore making it exceedingly difficult to predict the COVID period using only pre-COVID information. In the same way, we can conclude that this COVID period does not have useful information about the post-COVID future, and, in fact, we have experimentally verified that the quality of the models decreased with the introduction of the COVID period, thus leading to the decision to exclude this temporal section from the training set.

It is, in a certain way, clear that the prediction of the influx in emergency rooms can be useful for a more efficient management of hospital services, but there is visible value added at user level, in the sense that they will get better

and faster care [30]. To sustain this claim, we can observe the impact that the number of non-urgent patients has on the waiting time before being treated in the Emergency Department. In Fig. 2, for the RHAs with the highest daily affluence, we observe a positive, moderate to strong correlation between the number of non-urgent patients and the waiting time. This is an indicator, not entirely unexpected, that ED overcrowding of non urgent patients can lead to a substantial increase in the average waiting time for all patients, urgent or non urgent.

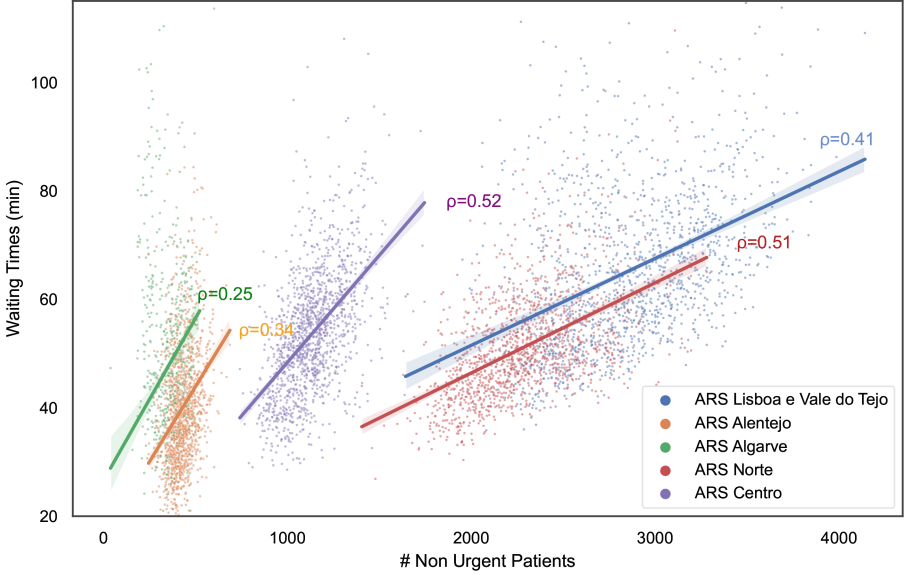


Fig. 2. Correlation between waiting times and non urgent patient volume. A strong to moderate correlation exist between these two variables, therefore implying that overcrowding increases waiting times.

4 Methods

4.1 Study Setting and Metrics

Now that we have presented the data used in this paper, let us define, and expose, the reasoning behind the rules by which we will create and evaluate the model.

- **Multivariate forecasting:** we want a model that leverages data and forecasts across different Regional Health Areas. In most research in the area, models are usually restricted to certain geographic areas, and a more general model, capable of working across different regions, might be able to uncover new interactions in data and increase robustness.

- **Long-time forecast:** In order to add value at the hospital management level, the forecast of the number of patients should not be limited to the following day or week. In this paper, we have chosen a 4-week (28-day) forecast, considering that it allows breathing room for management and personnel decisions. To the best of our knowledge, few works have worked on such an extended forecast horizon [6, 7], with only partial success.
- **Probability prediction:** besides obtaining an estimate of the most likely value in the future, a model that presents a probability density function on the prediction conveys much more information. Of special value is, for example, the definition of confidence intervals, which can transmit to those who use the model an idea of the confidence, or precision, of the model in its estimation. Almost all classical linear methods, such as ARIMA or Exponential Smoothing, are able to deliver confidence intervals over the predictions. However, the same is not true for common Neural Networks architectures.
- **Explanatory variables:** Importantly, we want to evaluate the predictive capacity of different variables, determining up to which passed time-step the model finds predictive value or which covariates, categorical or numerical, have a significant impact on the prediction.

The covariates that we intend to evaluate as explanatory variables are: day of the year, month and weekend, holidays, total number of patients in emergency rooms, number of unscheduled consultations in health centres, waiting time, patients with respiratory problems and total number of consultations in health centres. We do not expect that all these variables are relevant or necessary to solve the problem we present, however, they were used precisely to assess how the models would deal with redundant variables.

In this paper, we use four metrics to evaluate the models. The Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{5} \sum_{k=1}^5 \frac{1}{n} \sum_{i=1}^n |y_i^k - \hat{y}_i^k| \quad (1)$$

the Root Mean Squared Error (RMSE),

$$\text{RMSE} = \frac{1}{5} \sum_{k=1}^5 \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^k - \hat{y}_i^k)^2} \quad (2)$$

the Mean Absolute Percentage Error (MAPE),

$$\text{MAPE} = \frac{1}{5} \sum_{k=1}^5 \frac{1}{n} \sum_{i=1}^n \frac{|y_i^k - \hat{y}_i^k|}{|y_i^k|} \quad (3)$$

and the Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{5} \sum_{k=1}^5 \frac{1}{n} \sum_{i=1}^n (y_i^k - \hat{y}_i^k)^2. \quad (4)$$

Since we are evaluating the predictions over several groups (RHAs), the total error will be the average across RHAs. The most common metric across the literature for ED forecasting is the Mean Absolute Percentage Error (MAPE) [7, 11], however, when the true value is close to zero, this metric becomes unreliable. It also places a heavier penalty on negative errors (when the predicted value is higher than the true value) [27]. To overcome that, outliers values very close to zero are removed for this particular metric. To correctly evaluate the out-of-sample predictive capacity of the model, the dataset is divided into three subsets: train, validation, and test. The training set represents roughly 3.5 years, while the validation set and the test set have 10 weeks of data, each. The validation set is used to optimise hyper parameters and to identify overfitting during training, while the test set is unseen until the end and is only used to produce the final results. It contains the last 10 weeks available, from December 2021 to February 2022.

4.2 Models

The first and simpler method used for forecasting is the replication of the last k time-steps. This technique, which is used as Baseline in this paper, is also referred to as the naïve algorithm. By evaluating this model on the validation, the optimal value for k was estimated to be 7, thus representing the weekly periodicity in the data.

For comparison, other models commonly used in this area were also applied, namely AutoRegressive Integrated Moving Average (ARIMA) with a seasonal component [1, 12, 21, 23, 34], and its multivariate variant Vector AutoRegressive Integrated Moving Average (VARIMA) [23].

Also used was the exponential Smoothing algorithm, a simple method that has also shown good results in the literature [8]. Finally, to gauge the performance of common machine learning models, the XGBoost model was used. Out of these models, the XGBoost [24] (a Decision Tree Boosting algorithm), is the only model capable of using past and future covariates, with the disadvantage of not being specifically tailored for time-series data.

The model used in this paper, however, is the Temporal Fusion Transformer (TFT). We chose this model because it achieves all the goals mentioned previously. To define the model input, we first need to separate variables into static, target and time dependent. Static covariates, such as time-series variance or mean, are specific to each group, i.e. RHA, and are defined as s_i with $i = 0, \dots, 4$. $y_{i,t}$ is the target for group i at time-step t and $x_{i,t} = [p_{i,t}^T, f_{i,t}^T]^T$ the time dependent covariates, with p representing past covariates, meaning covariates that are only known until the present, as f future covariates, that can be assumed to be known in the past and the future, in our case, holidays and weekends.

Table 1. Hyperparameters for TFT model after tuning.

Hyperparameter	Value
Encoder length	42
Batch size	40
Prediction length	28
Gradient clipping	0.022730
Learning rate	0.0011149
Hidden size	33
Number of attention heads	8
Dropout	0.19230
Hidden continuous size	19

The prediction function is defined as [26]:

$$\hat{y}_i(q, t, \tau) = f_q(\tau, y_{i,t-k:t}, x_{i,t-k:t}, s_i) \quad (5)$$

where $\hat{y}_i(q, t, \tau)$ is the predicted q th quantile for the $\tau \in \{1, \dots, \tau_{max}\}$ value in group i , at time t . For the specific case of this work, $\tau_{max} = 28$, as we want to forecast simultaneously 28 days ahead. By predicting quantiles, we obtain a quasi-distribution of the expected value, and gain the capacity to define confidence intervals.

Initially introduced by [26], this model instantiated a novel architecture, combining a few mechanisms previously only used separately, in a single model. The key features of the TFT are:

- Variable Selection Network: three independent Selection Networks, one for each variable set, to select only relevant variables at each time-step. This module removes noisy variables that do not add predictive value, while giving some level of insight into the variables that are more significant to the prediction;
- A Gating Mechanism to skip any other element of the architecture. For specific cases where exogenous variables are not useful or there is no need for non-linear processing (e.g. in very simple forecasts) the Gating Mechanism, also referred to as Gated Residual Network [16], allows the model to only use non-linear processing when needed;
- Static Variables encoding to combine static information with time-series data;
- Temporal Dependency Processing to capture short-term dependency, with an LSTM encoder-decoder [13, 18], and long-term dependency using a Multi-Head Attention mechanism [38]. By an additive aggregation of the different heads, this mechanism gains explainability, as the weights in the aggregated Multi-head represent time-step importance;
- Confidence Intervals: the output of the models are quantiles, that define prediction intervals, at each forecast time-step.

To obtain the quantile predictions, a specific loss, the Quantile Loss, is defined as [25]:

$$QL(y, \hat{y}, q) = \max\{q(y - \hat{y}), (q - 1)(y - \hat{y})\} \quad (6)$$

for each quantile q . The final Loss is the average QL across quantiles and for the entire prediction horizon $[0, \tau_{max}]$. In this work, the quantiles used were $[0.02, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98]$. When $q = 0.5$ the Loss is equal to MAE divided by 2, and $q = 0.5$ (the median) is the value used for the point-wise prediction of the model.

The overall architecture of the TFT can be seen in Fig. 3 and the hyperparameters are defined in Table 1.

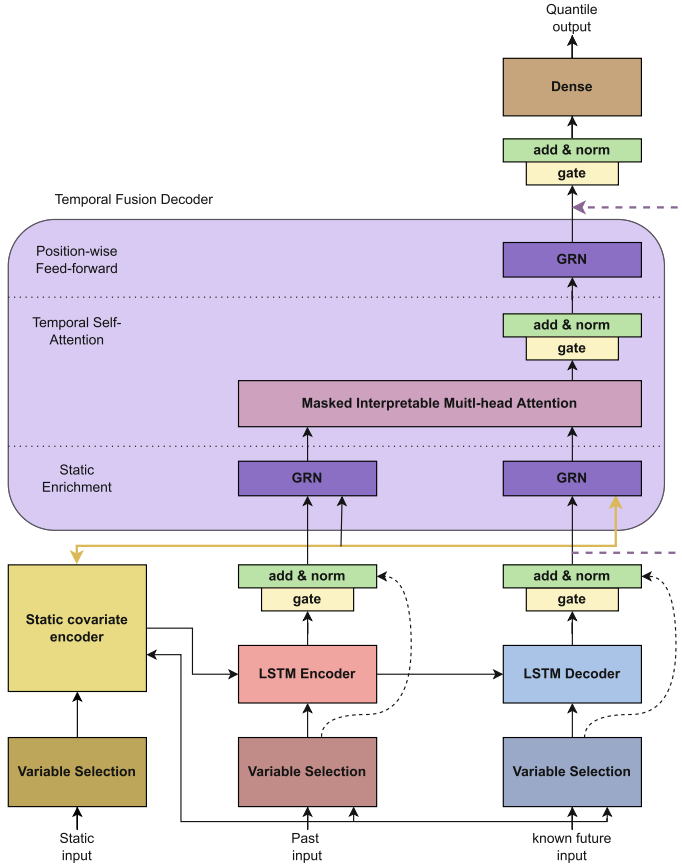


Fig. 3. TFT architecture. The inputs are static metadata, time-varying past inputs (including past target values) and known future information. The Variable selection unit selects the most relevant features, while the Gated Residual Network allows to skip over unused sections of the architecture. The interpretable multi-head attention is used to evaluate the most relevant time-steps. Image adapted from [26].

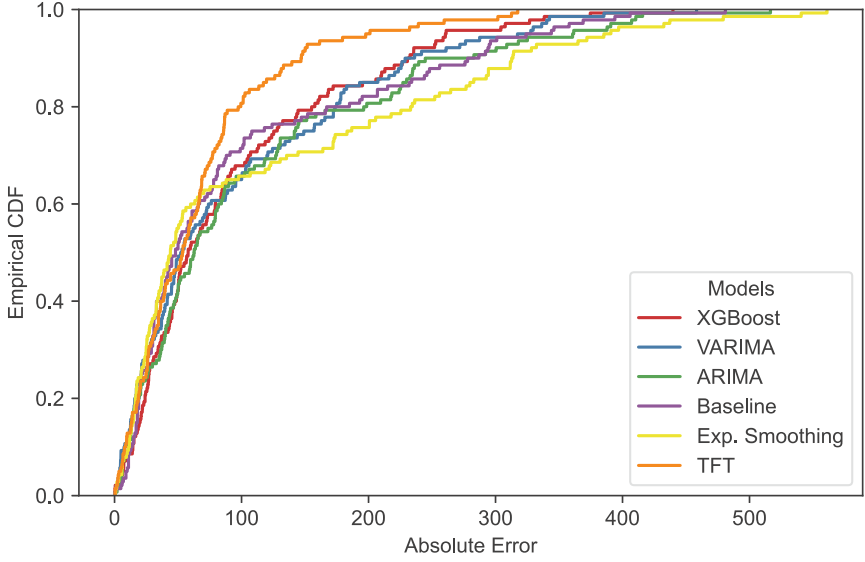
Table 2. Prediction accuracy for various models in the period 24/01/2022–20/02/2022. To evaluate the models, four metrics are used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE). Bold indicates the best result; TFT is consistently more accurate than the baselines.

Models	MAE	RMSE	MAPE	MSE
Baseline	95.1643	116.5850	7.3483	20245.0643
Exp. Smoothing	112.5885	135.6158	7.3135	29888.6468
ARIMA	104.9886	129.6084	7.8484	22949.7471
VARIMA	94.6441	120.6674	7.9250	18407.9554
XGBoost	92.0307	112.3295	7.7027	16178.5531
TFT	66.7551	84.4102	5.9084	8379.7340

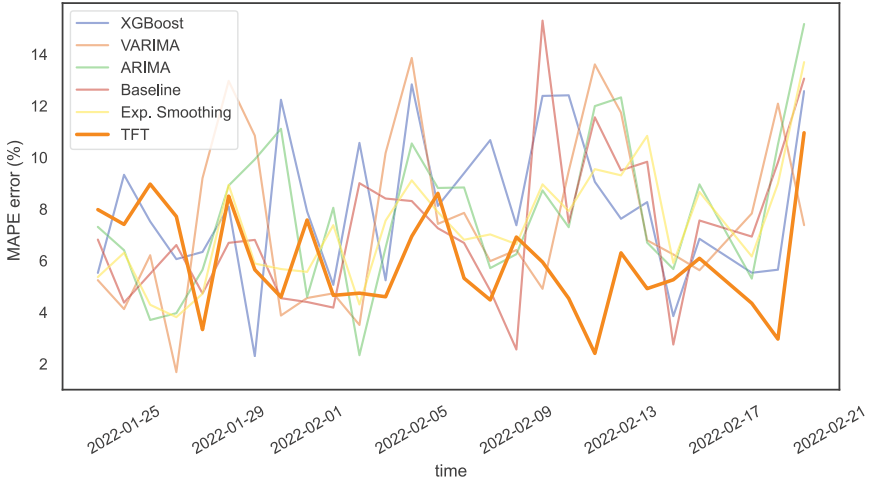
5 Results

In this section, we present the results of the TFT and the other models for a 4 week forecast window. Table 2 illustrates how the TFT outperforms other common models in the literature for long time prediction, with a Root Mean Squared Error (RMSE) of 84.4102, or approximately 84 people per day. This metric, however, might be deceptive, as it is scale dependent, meaning that RHAs with a larger daily volume will necessarily yield a higher RMSE, and skew the results. The Mean Absolute Percentage Error (MAPE) on the other hand, is scale invariant, and it better depicts the overall predictive power of the models, with the TFT obtaining a 5.91% percentage error. Taking a more detailed look at the predictions, in Fig. 6, we can see how the model can make predictions at different scales, correctly representing two characteristics that we know are part of the data, the weekly cycle, and the peak of users on Monday. To better compare the models, we utilised an empirical CDF for each model, as seen in Fig. 4a. In this Figure, depicting Absolute Error, the TFT shows overall better performance. We also acknowledge that the Exponential Smoothing algorithm obtains favourable results for roughly half of the predictions. As suggested in Fig. 4b, the TFT outperforms the other models in the last 2 weeks of the forecast window. This illustrates the superior capability of deep learning models to perform long term prediction, as the complexity of the model helps identify long term patterns.

But the strength of the Temporal Fusion Transformer used goes beyond the precision of the model. First, we can observe the attention given to each time-step. As explained in Sect. 4.2, attention is used to identify which input elements, containing up to 6 weeks of data, are most useful during forecast. In Fig. 5, it can be distinguished how the model values the most recent time-steps with a higher weight, which is intuitively expected and shows that old information has less value to the model. This validates a common assumption in linear models, that ascribe more weight to more recent observations, as is the case of the Exponential Smoothing model. In Fig. 6 we can also verify this effect, with the grey line over



(a) Empirical Cumulative Distribution function for the absolute difference between the true value and the predicted value, for all RHA.



(b) Mean Absolute Percentage Error (MAPE) over time. Most models have increased error in the last two weeks, with the exception of the TFT model.

Fig. 4. Comparative analysis of model prediction.

the input period representing attention. In the forecasting figures, we can observe that different RHAs have different attention weights depending on the input

vector of the model. In addition, we observe another more intriguing feature, which are spikes in attention during the weekend, this may happen because particular attention is given to one or two previous weekends to define patient volume in future weekends.

After having determined that the model attributes higher attention to more recent time-steps, we will now observe the importance attributed by the model to the covariates. We can categorise covariates into three categories: static, past, and future. In Fig. 7, it is possible to observe the importance attributed to each past or future covariate. In the left side figure, we see that the variable with the most weight is the percentage of patients in the emergency room with respiratory problems. For this period, excess affluence in hospital emergency rooms could be attributed to peaks in influenza/COVID-19 transmission, it therefore makes sense that this variable can be a predictive indicator of future positive trends in the number of non-urgent cases. The second most important variable is patient waiting time, which is in line with the positive relationship presented at the beginning of this article between the increase in waiting time and the increase in non-urgent patients. However, we should not focus our attention solely on the variables relevant to the model. There is interest in observing the variables that did not add value to the model; here we can observe that the information regarding health care centres (*n.cons.total,prog*) did not add value to the model, meaning that there is no clear interaction between patient volume in health care centres, mostly used for primary health care and minor health issues, and non-urgent patients in Emergency Departments.

As we see on the right side of Fig. 7, the number of known covariates in the future is a smaller part of the total number of covariates. The most important feature is the categorical variable indicating public holidays in Portugal. The model has attributed such an importance to holidays because they have a

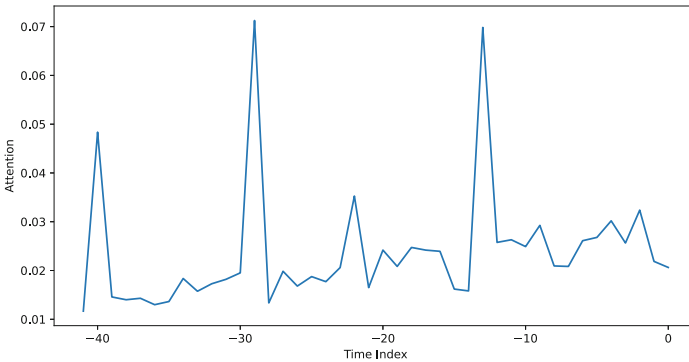


Fig. 5. Average attention attributed over the input vector. More recent time-steps are given more value than older time-steps.

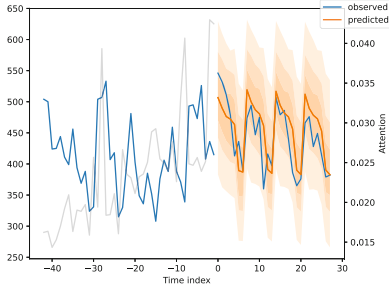
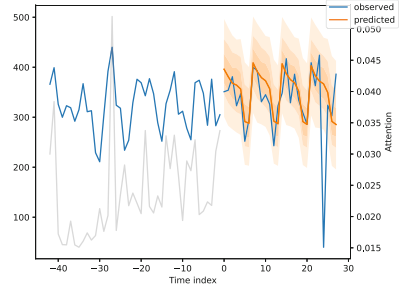
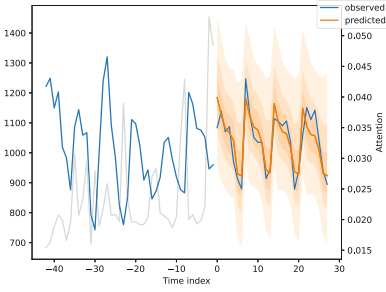
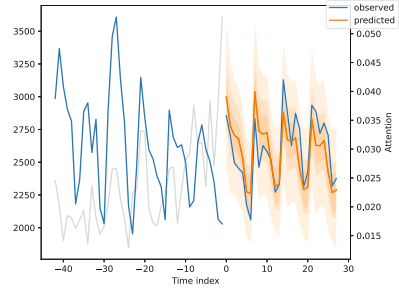
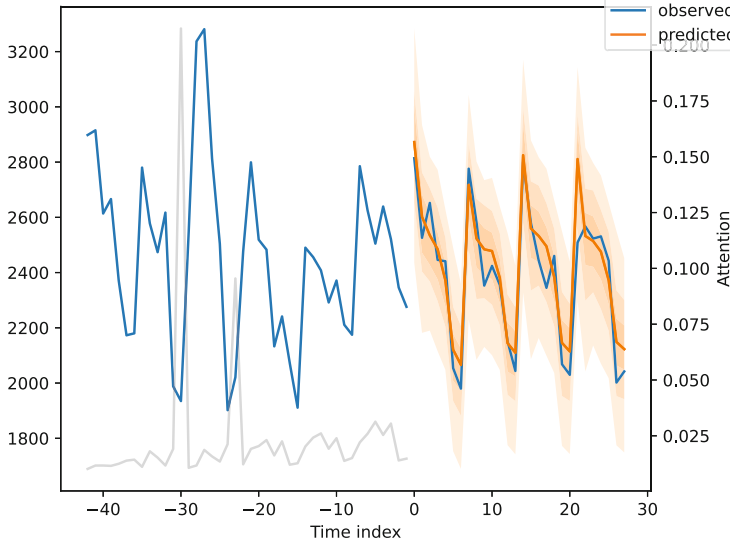
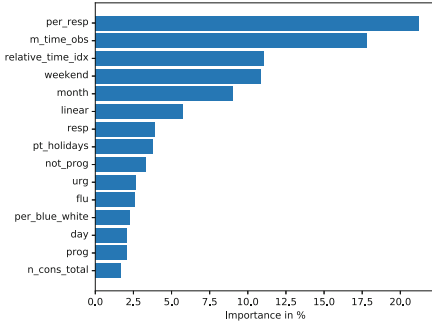
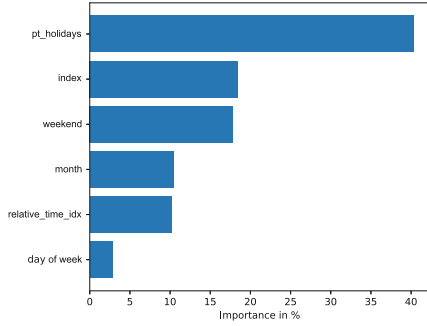
(a) *ARS Alentejo*(b) *ARS Algarve*(c) *ARS Centro*(d) *ARS Lisboa e Vale do Tejo*(e) *ARS Norte*

Fig. 6. Predictions over the test set. Over the input vector, we can see the grey line representing attention. In orange is the median predictive value ($q = 0.5$), with different quantiles shown as shaded area (Color figure online).



(a) Past Covariates Importance.



(b) Future Covariates Importance.

Fig. 7. Variable Importance. The most relevant past feature to the model is the percentage of patients in ED with indication of respiratory problems. For future covariates, variables that are known in the future, the most relevant is a feature that indicates public holidays in Portugal.

severe impact on patient volume, not only on the day, but also on the next day, when close to the weekend. Furthermore, the other future covariates have a non-negligible importance both as past and future covariates, thus supporting the claim found in the literature that calendar variables have a significant impact on the prediction.

6 Conclusion

This paper presented a novel application of the Temporal Fusion Transformer (TFT) model to predict non-urgent patient volume in Portuguese public hospitals by Health Regional Areas (HRA). The results were encouraging, surpassing other models commonly found in the literature [21,23]. The forecasting of an entire month is seldom done in the literature [2,7], and the model presented did not show signs of deterioration over the forecast window; despite that, it would be interesting to drive the forecasting period even further, either by autoregression or by increasing the forecast window, so as to analyse the maximum prediction length of the model, or a potential trade-off between forecast window and predictive quality.

The introduction of a multivariate model with good results across groups is a positive prospect, since one limitation of univariate time-series is the natural low-data regimen, while multivariate models can merge information from multiple sources, thus increasing the total amount of data fed to the models. In the future, this model can increase in granularity, forecasting at the hospital level instead of aggregated values by HRAs. Although a greater challenge, due to the increased noise and randomness that comes from the decrease in the study population, we expect that the combination of a large number of time-series could improve the robustness and global quality of the model, specifically if we add more relevant

static variables. For this paper, only HRA and time-series statistics were used as static covariates, but as noted in [14], across different regions there is distinct demand for emergency care, thus impacting the scale and variance of the time-series. In future work, we plan to introduce other factors that might contribute to encode region-specific information as static covariates, such as demographics, modes of transport available, socio-economic characterisation of the patient population and number and capacity of private health care providers in the region. All these elements might help to represent each class, and ultimately be used for a generalisation of the model to unseen hospitals, where these variables might help to represent how similar a new unseen hospital/RHA is to hospitals/RHAs in the training data.

References

1. Abdel-Aal, R., Mangoud, A.: Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis. *Comput. Meth. Programs Biomed.* **56**(3), 235–247 (1998). [https://doi.org/10.1016/s0169-2607\(98\)00032-7](https://doi.org/10.1016/s0169-2607(98)00032-7)
2. Aboagye-Sarfo, P., Mai, Q., Sanfilippo, F.M., Preen, D.B., Stewart, L.M., Fatovich, D.M.: 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 (2015). <https://doi.org/10.1016/j.jbi.2015.06.022>
3. Afilal, M., Yalaoui, F., Dugardin, F., Amodeo, L., Laplanche, D., Blua, P.: Forecasting the emergency department patients flow. *J. Med. Syst.* **40**(7), 1–18 (2016). <https://doi.org/10.1007/s10916-016-0527-0>
4. Batal, H., Tench, J., McMillan, S., Adams, J., Mehler, P.S.: Predicting patient visits to an urgent care clinic using calendar variables. *Acad. Emerg. Med.* **8**(1), 48–53 (2001). <https://doi.org/10.1111/j.1553-2712.2001.tb00550.x>
5. Bernstein, S.L., et al.: The effect of emergency department crowding on clinically oriented outcomes. *Acad. Emerg. Med.* **16**(1), 1–10 (2009). <https://doi.org/10.1111/j.1553-2712.2008.00295.x>
6. Boyle, J., et al.: Predicting emergency department admissions. *Emerg. Med. J.* **29**(5), 358–365 (2012). <https://doi.org/10.1136/emj.2010.103531>
7. Carvalho-Silva, M., Monteiro, M.T.T., de Sá-Soares, F., Dória-Nóbrega, S.: Assessment of forecasting models for patients arrival at emergency department. *Oper. Res. Health Care* **18**, 112–118 (2018). <https://doi.org/10.1016/j.orhc.2017.05.001>
8. Champion, R., et al.: Forecasting emergency department presentations. *Aust. Health Rev.* **31**(1), 83–90 (2007). <https://doi.org/10.1071/AH070083>
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). <https://doi.org/10.18653/v1/N19-1423>
10. Diehl, A.K., Morris, M.D., Mannis, S.A.: Use of calendar and weather data to predict walk-in attendance. *South. Med. J.* **74**(6), 709–712 (1981). <https://doi.org/10.1097/00007611-198106000-00020>

11. Ekström, A., Kurland, L., Farrokhnia, N., Castrén, M., Nordberg, M.: Forecasting emergency department visits using internet data. *Ann. Emerg. Med.* **65**(4), 436–442.e1 (2015). <https://doi.org/10.1016/j.annemergmed.2014.10.008>
12. Eyles, E., Redaniel, M.T., Jones, T., Prat, M., Keen, T.: Can we accurately forecast non-elective bed occupancy and admissions in the NHS? A time-series MSARIMA analysis of longitudinal data from an NHS trust. *BMJ Open* **12**(4) (2022). <https://doi.org/10.1136/bmjopen-2021-056523>
13. Fan, C., et al.: Multi-horizon time series forecasting with temporal attention learning. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2019*, pp. 2527–2535. Association for Computing Machinery, New York (2019). <https://doi.org/10.1145/3292500.3330662>
14. Farmer, R.D., Emami, J.: Models for forecasting hospital bed requirements in the acute sector. *J. Epidemiol. Commun. Health* **44**(4), 307–312 (1990). <https://doi.org/10.1136/jech.44.4.307>
15. Harrou, F., Dairi, A., Kadri, F., Sun, Y.: Forecasting emergency department overcrowding: a deep learning framework. *Chaos, Solitons Fractals* **139**, 110247 (2020). <https://doi.org/10.1016/j.chaos.2020.110247>
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
17. Hertzum, M.: Forecasting hourly patient visits in the emergency department to counteract crowding. *Ergon. Open J.* **10**(1) (2017). <https://doi.org/10.2174/1875934301710010001>
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
19. Holleman, D.R., Bowling, R.L., Gathy, C.: Predicting daily visits to a walk-in clinic and emergency department using calendar and weather data. *J. Gen. Intern. Med.* **11**(4), 237–239 (1996)
20. Hurwitz, J.E., Lee, J.A., Lopiano, K.K., McKinley, S.A., Keesling, J., Tyndall, J.A.: A flexible simulation platform to quantify and manage emergency department crowding. *BMC Med. Inform. Decis. Mak.* **14**(1), 50 (2014). <https://doi.org/10.1186/1472-6947-14-50>
21. Jones, S.S., Thomas, A., Evans, R.S., Welch, S.J., Haug, P.J., Snow, G.L.: Forecasting daily patient volumes in the emergency department. *Acad. Emerg. Med.* **15**(2), 159–170 (2008). <https://doi.org/10.1111/j.1553-2712.2007.00032.x>
22. Kadri, F., Abdennbi, K.: RNN-based deep-learning approach to forecasting hospital system demands: application to an emergency department. *Int. J. Data Sci.* **5**, 1–25 (2020). <https://doi.org/10.1504/IJDS.2020.10031621>
23. Kadri, F., Harrou, F., Chaabane, S., Tahon, C.: Time series modelling and forecasting of emergency department overcrowding. *J. Med. Syst.* **38**(9), 1–20 (2014). <https://doi.org/10.1007/s10916-014-0107-0>
24. Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems, NIPS 2017*, vol. 30, pp. 3149–3157. Curran Associates Inc., Red Hook, NY, USA (2017). <https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>
25. Koenker, R., Hallock, K.F.: Quantile regression. *J. Econ. Perspect.* **15**(4), 143–156 (2001). <https://doi.org/10.1257/jep.15.4.143>
26. Lim, B., Arık, S.O., Loeff, N., Pfister, T.: Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **37**(4), 1748–1764 (2021). <https://doi.org/10.1016/j.ijforecast.2021.03.012>

27. Makridakis, S.: Accuracy measures: theoretical and practical concerns. *Int. J. Forecast.* **9**(4), 527–529 (1993). [https://doi.org/10.1016/0169-2070\(93\)90079-3](https://doi.org/10.1016/0169-2070(93)90079-3)
28. Milner, P.: Forecasting the demand on accident and emergency departments in health districts in the trent region. *Stat. Med.* **7**(10), 1061–1072 (1988). <https://doi.org/10.1002/sim.4780071007>
29. Navares, R., Díaz, J., Linares, C., Aznarte, J.L.: Comparing ARIMA and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid. *Stoch. Env. Res. Risk Assess.* **32**(10), 2849–2859 (2018). <https://doi.org/10.1007/s00477-018-1519-z>
30. Pines, J.M., Hollander, J.E.: Emergency department crowding is associated with poor care for patients with severe pain. *Ann. Emerg. Med.* **51**(1), 1–5 (2008). <https://doi.org/10.1016/j.annemergmed.2007.07.008>
31. Pulkkinen, E.: forecasting emergency department arrivals with neural networks. Bachelor's thesis, Tampere University, Tampere, Finland (2020)
32. Rathlev, N.K., et al.: Time series analysis of variables associated with daily mean emergency department length of stay. *Ann. Emerg. Med.* **49**(3), 265–271 (2007). <https://doi.org/10.1016/j.annemergmed.2006.11.007>
33. Rocha, C.N., Rodrigues, F.: Forecasting emergency department admissions. *J. Intell. Inf. Syst.* **56**(3), 509–528 (2021). <https://doi.org/10.1007/s10844-021-00638-9>
34. Schweigler, L.M., Desmond, J.S., McCarthy, M.L., Bukowski, K.J., Ionides, E.L., Younger, J.G.: Forecasting models of emergency department crowding. *Acad. Emerg. Med.* **16**(4), 301–308 (2009). <https://doi.org/10.1111/j.1553-2712.2009.00356.x>
35. Sharafat, A.R., Bayati, M.: PatientFlowNet: a deep learning approach to patient flow prediction in emergency departments. *IEEE Access* **9**, 45552–45561 (2021). <https://doi.org/10.1109/ACCESS.2021.3066164>
36. Sudarshan, V.K., Brabrand, M., Range, T.M., Wiil, U.K.: Performance evaluation of emergency department patient arrivals forecasting models by including meteorological and calendar information: a comparative study. *Comput. Biol. Med.* **135**, 104541 (2021). <https://doi.org/10.1016/j.combiomed.2021.104541>
37. Tuominen, J., et al.: Forecasting daily emergency department arrivals using high-dimensional multivariate data: a feature selection approach. *BMC Med. Inform. Decis. Mak.* **22**, 134 (2022). <https://doi.org/10.1186/s12911-022-01878-7>
38. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc. (2017). <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
39. Wargon, M., Guidet, B., Hoang, T.D., Hejblum, G.: A systematic review of models for forecasting the number of emergency department visits. *Emerg. Med. J.* **26**(6), 395–399 (2009). <https://doi.org/10.1136/emj.2008.062380>
40. Whitt, W., Zhang, X.: Forecasting arrivals and occupancy levels in an emergency department. *Oper. Res. Health Care* **21**, 1–18 (2019). <https://doi.org/10.1016/j.orhc.2019.01.002>
41. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics (2020). <https://doi.org/10.18653/v1/2020.emnlp-demos.6>

42. Zachariasse, J.M., van der Hagen, V., Seiger, N., Mackway-Jones, K., van Veen, M., Moll, H.A.: Performance of triage systems in emergency care: a systematic review and meta-analysis. *Br. Med. J. Open* **9**(5) (2019). <https://doi.org/10.1136/bmjopen-2018-026471>
43. Zhou, L., Zhao, P., Wu, D., Cheng, C., Huang, H.: Time series model for forecasting the number of new admission inpatients. *BMC Med. Inform. Decis. Mak.* **18**(1), 39 (2018). <https://doi.org/10.1186/s12911-018-0616-8>