Anomalies and breakpoint detection for a dataset of firefighters' operations during the COVID-19 period in France

Roxane Elias Mallouhy[1], Christophe Guyeux[2], Chady Abou Jaoude[3], and Abdallah Makhoul[2]

[1] Prince Mohammad Bin Fahd University, Khobar, Kingdom of Saudi Arabia.
reliasmallouhy@pmu.edu.sa,
[2] University of Bourgogne Franche-Comté, Belfort,France.
christophe.guyeux@univ-fcomte.fr,
abdallah.makhoul@univ-fcomte.fr,
[3] Antonine University, Baabda, Lebanon. chady.aboujaoude@ua.edu.lb

Abstract. Firefighters are exposed to many hazards. The main objective of this study is to apply machine learning techniques to tailor the need for firemen operations to their demands. This strategy enables fire departments to organize their resources, which leads to a reduction of human, material and financial requirements. This work focuses on predicting the number of firefighters' interventions during the sensitive period of the global pandemic COVID-19. Experiments applied to a dataset from 2016 to 2021 provided by the Fire and Rescue Department, SDIS 25, in the region Doubs-France have shown an accurate prediction and revealed the existence of a turning point in August 2020 due to an increase in coronavirus cases in France.

Keywords: prediction, firefighters, feature selection, breakpoint, anomalies detection, COVID-19

1 Introduction

The World Health Organization (WHO) recorded the first pneumonia of unknown cause on December 31, 2019, in Wuhan, China. A few days later, on January 24, 2020, it was confirmed that the virus had reached France. Countries could never have imagined that such a devastating viral disease would emerge and a pandemic can cause morbidity around the world. The ability of public health actors, populations, and institutions to prepare for and respond effectively to such crisis in order to maintain essential functions was not satisfied worldwide. Hospitals faced bed shortages, delayed non-urgent surgeries, and mobilised all medical staff resources. This inevitably made the fire and rescue services (SIS) in France the first actors of aid and emergency care providers in the operational response to this epidemic. They are the key players in this crisis, in caring for people and in supporting the health system by participating in

transports between hospitals by land or by helicopter. They also run facilities for the elderly ('Etablissement d'Hébergement Pour Personnes Agées Dépendantes EHPAD'): examination, disinfection, distribution of meals, etc. Therefore, the use of machine learning is very important in such a case: predicting the number of firemen interventions can directly lead to a reduction in financial, material, and human resources. Consequently, the efficiency of emergency operations during this pandemic will be improved.

The database for this study illustrated in Figure 1 was provided by the Fire and Rescue Department, SDIS 25, in the Doubs-France region. It contains hourly recorded operations from January 2015 to June 2021. The aim of this work is to predict the number of firefighters by performing several steps, starting with the reduction of the number of features, the detection of the breakpoint related to the coronavirus period, the comparison of the selected features before and after the COVID-19 and finally the detection of the anomalies.

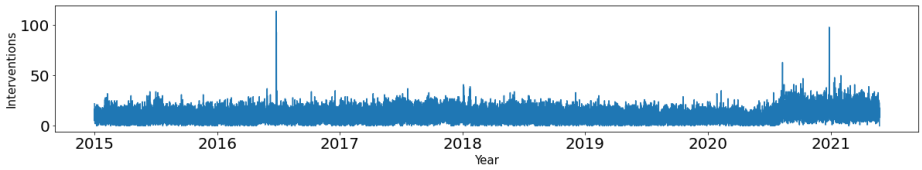


Fig. 1: Dataset presentation

This paper is arranged as follows: Section 2 gives a brief description of the contributions of related works, Section 3.1 presents the feature selection method used to reduce the number of attributes of the database, Section 3.2 describes how the breakpoint was selected and shows its relationship with the selected features and Section 3.3 shows the anomaly detection. In the following two sections, we present and discuss the obtained results and finally in Section 6 we draw a brief conclusion.

2 State of the art

Machine learning and artificial intelligence have played an important role in research, healthcare, and even agriculture during the COVID-19 period. ML Technology allows computers to interpret large amounts of data to quickly identify patterns and insights and understand the pathobiology of this disease. Numerous studies have been proposed in various countries to predict the spread of COVID-19. In India, a study was conducted using linear regression, multilayer perceptron and vector autoregression to predict the spread of this disease [6]. Likewise, a hybrid machine learning approach based on adaptive network-based fuzzy inference system and multilayer perceptron imperialist-competitive algorithm was proposed to predict the individuals infected by coronavirus and mortality in Hungary [7]. Furthermore, Machine Learning helped detect COVID-19

at an early stage, which helped monitor disease progression and potentially reduce mortality. One study was validated on a dataset of chest X-rays and CT scan compared popular deep learning-based feature extraction frameworks for automatic COVID-19 classification by sorting subjects as control or COVID-19 [8]. In addition, Artificial Intelligence played an important role in combating the growing pandemic. If enough data is trained by a deep learning model, it can help in finding an effective vaccine candidate by detecting patterns in the data [9]. A research paper proposed a silico approach for the prediction and design of a multi-epitope vaccine (DeepVacPred), which helps speed up the process of vaccine development. The applied framework predicts 26 vaccine subunits from the existing SARS-CoV-2 spike protein sequence and has proven to cope with recent mutations of the virus [10].

On the other hand, we found works related to the same dataset used in our study provided by SDIS 25 such as [1], in which a machine learning based method was proposed to predict the turnaround time of each ambulance at a specific time and hospital. Temporal and external variables were considered and irregularly spaced time series model were utilized. Similarly, various artificial intelligence techniques were implemented to predict the number of firefighters' interventions such as Auto Regression, Moving Average, Auto Regressive Integrated Moving Average and Prophet [5], Long Short Term Memory [2], Extreme Gradient Boosting [3] and neural network accurately Multi-Layer Perceptron [4]. To the authors' knowledge, there is no research work on the specific task of using features, breakpoints, and anomalies for dataset analysis to predict the number of firemen during COVID-19.

3 Methodology

In this section, the methodology used for prediction the number of firefighters' interventions is explained. First, the number of attributes in the dataset was reduced by applying a feature selection technique, then breakpoints were detected, and finally anomalies were found and replaced. At each step, statistical features, specifically mean absolute error and root mean square error, were calculated to evaluate the method used. However, since the choice of hyperparameters directly reflects the performance of the model, Optuna [11], a hyperparameter optimization system, was used to optimize the parameters of the XGboost algorithm [12], specifically learning_rate, max_depth, random_state, n_estimators, and n_jobs. All machine learning models in this research were run using Jupyter Notebook on a 2.7 GHz Core i7 processor equipped by 8 GB RAM.

3.1 Feature Selection

The presence of 1573 attributes certainly complicates the processing of the dataset. Therefore, feature selection is required before we proceed to the next stage of this study. The idea is to retain only useful features by removing irrelevant attributes since adding more and more variables to a dataset increases the

overall complexity and could reduce the accuracy of a classifier. Thus, a feature is considered irrelevant if it has been removed without affecting the performance of the model. The feature selection approach used in this paper is called 'feature importance' which is a built-in class, where a score is assigned to each feature in the dataset, and the higher the score, the more relevant the feature is [18]. Three machine learning algorithms were implemented: XGBOOST [12], Extra Tree Classifier [13] and Random Forest Regressor [14]. For each algorithm, feature importance was calculated and the common best top 16 attributes were extracted including the 'lockdown' feature corresponding to the period of COVID-19.

3.2 Breakpoint detection

In the dataset used (Figure 1), it is quite evident to recognize an increase in the number of firemen interventions starting August 2020 which is closely related to the augmentation in the number of COVID-19 cases in France as can be seen in Figure 2, which shows the number of active cases since the emergence of coronavirus disease in France until today according to Worldometer statistics. Thus, the objective of change point detection is to highlight the direct impact of coronavirus on the number of firefighters' interventions. For this purpose, a Python library for the analysis of breakpoints called 'rupture' was applied [15]. After running the code, a change point was detected on August 5, 2020 presented in a green dotted line in Figure 3. Therefore, we divided the dataset into two periods separated by the breakpoint: Pre and post COVID-19 spread.

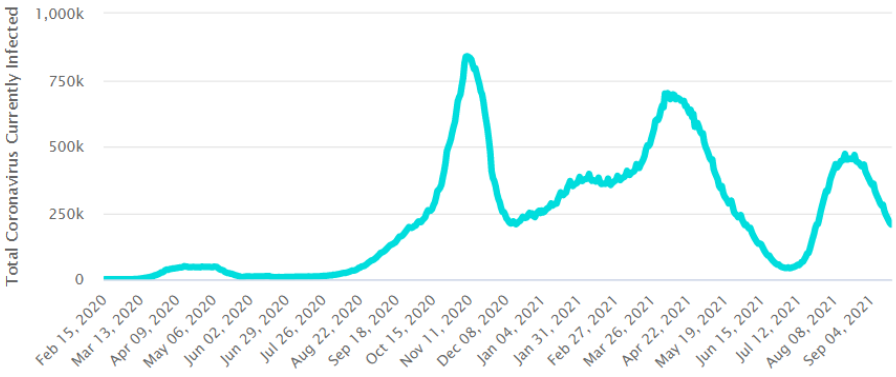


Fig. 2: COVID-19 active cases in France

Further action was taken by rechecking the order of feature importance to ensure that the attributes related to the COVID-19 differed if they were before or after the breakpoint found. Table 1 shows that the score of the three features 'confinement1', 'confinement2' and 'couvrefeux' (representing 'lockdown1', 'lockdown2' and 'curfew', respectively) associated with coronavirus disease change

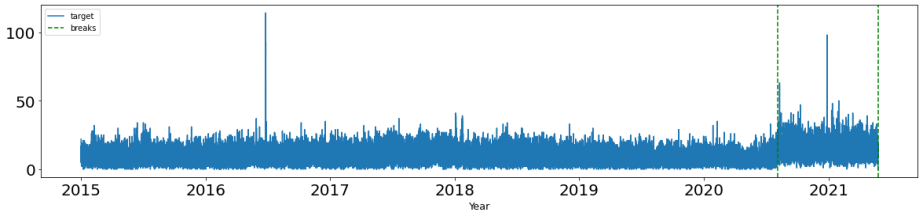


Fig. 3: Breakpoint detection

immediately before and after the identified breakpoint. Indeed, the French president announced a nationwide lockdown starting March 17, 2020 until May 10. A second lockdown then occurred on October 30, 2020 until mid-December. However, the curfew will remain in place from December 15, 2020 to June 20, 2021. While the goal of the lockdown and curfew may seem similar, there is a major difference between the two: During curfew, people are forced to stay at home for a specific number of hours and major services remain closed for a certain period of time.

Period	Before 5/8/2020	After 5/8/2020
Feature	Feature importance score	
confinement1	0.024589	0.000000
confinement 2	0.000000	0.018046
couvrefeux	0.000000	0.056813

Table 1: Feature Importance Before and After COVID-19 peak period

3.3 Anomalies detection

Anomaly detection is required in such huge metadata to find elements that trigger suspicions that might be different from the majority of the data. Logically, these anomalies can be associated with rare events such as water floods, large fires, power outages, and of course COVID 19 disease. In this work, Isolation Forest, an advanced statistical learning technique, was used [16] by importing PyCaret library [17]. This method assumes that some data points are more dominant than others, making outliers susceptible to the isolation mechanism. 1122 anomalous data points are presented as red dots in Figure 4. It can be clearly seen that most of the outliers are detected in the time after the breakpoint founded in Section 3.2.

4 Number of firefighters' interventions

After applying all the above steps, two experiments were performed on the dataset. The first experiment aims to predict the number of firemen interven-

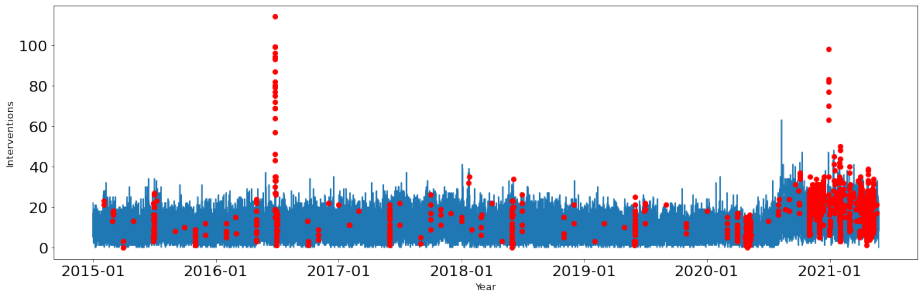


Fig. 4: Anomalies detection

tions for any period or date. This requires replacing the detected anomalies with the mean of the non-outlier values, as shown in the Figure 5, then the breakpoint (presented in green dots) was recalculated in Figure 6 to test if the changes influenced the selection of the coronavirus period as the point of change in the dataset, and finally the interventions for the period before and after the peak are predicted in Figure 7.

On the other hand, the second experiment predict the values considered abnormal as depicted in Figure 8 by replacing the normal data points with their mean. The idea behind this is to test the efficiency of our model in predicting an abnormal value that is certain to deviate from the normal trend particularly during the coronavirus pandemic.

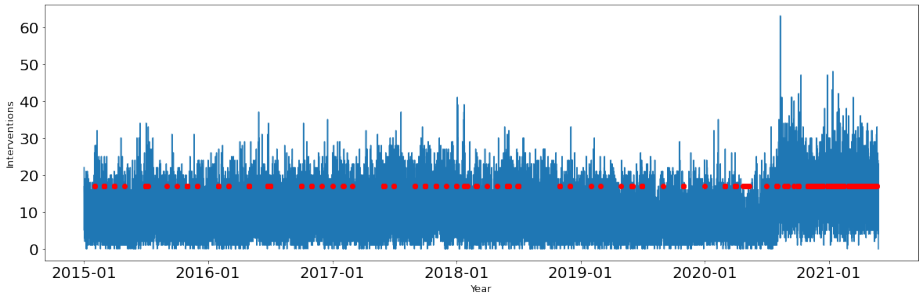


Fig. 5: Dataset after replacing the anomalies

In addition, to check how accurate the prediction is for the test set, the period before August 5, 2020, and the period after August 5, 2020, the Mean Absolute Error and Root Mean Square Error metrics were used as indicated in Figure 9. Technically, these statistical characteristics show the error between the actual and predicted values.

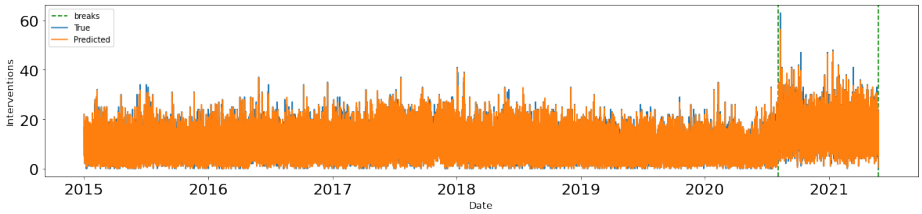
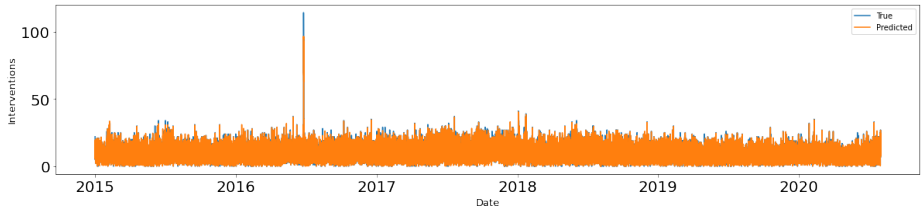
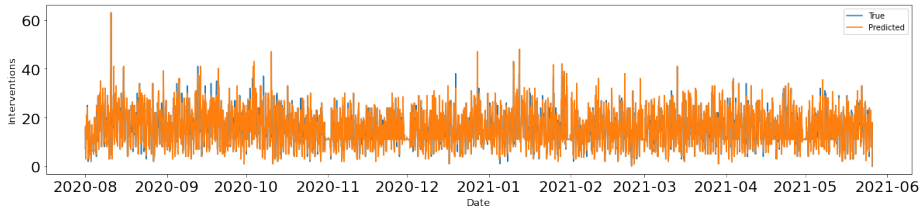


Fig. 6: Breakpoint detection and interventions prediction after replacing the anomalies



((a)) Period pre COVID-19 peak



((b)) Period pre COVID-19 peak

Fig. 7: Breakpoint detection after experiment 1

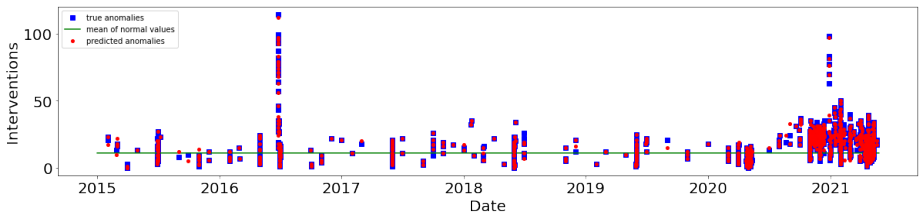


Fig. 8: Real vs prediction for anomalies values

5 Discussion

Feature selection in this study was a turning point to overcome the challenge of a dataset with 1572 columns and to bring down the computational time by

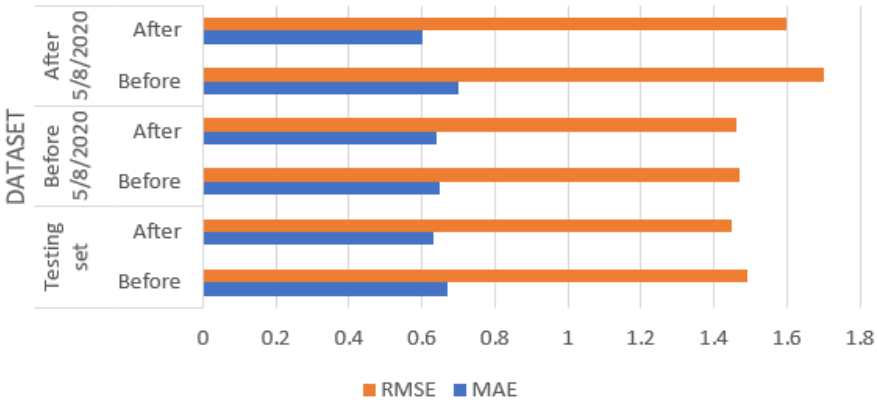


Fig. 9: Statistical Features for the whole dataset, period pre and post COVID-19 peak

eliminating irrelevant attributes. This technique gets significant results as the root mean square error decreased from 2.192 to 1.49 and on the other hand, the training time was reduced from 230.382 seconds to 19.419 seconds.

Besides, it is highly suggestive that the breakpoint discovered on August 5, 2020 is directly related to the rise in COVID-19 cases in France, which has led to a significant increase in the number of firefighters' interventions. Nearly three months after the initial lockdown, this peak was reached when the virus began to attack the younger generation. In this step of the analysis, the feature selection method was replicated for the periods before and after the breakpoint. As the results show, the importance scores of 'confinement2' and 'couvrefeux' were zero for the dataset before the peak of COVID-19 and increased thereafter. However, for 'confinement 1' the values were reversed. Here we see that the real pandemic crisis in France starts as early as August 2020: the number of new cases reached an average of 3,003 per day, a figure four times higher than the average of 746 per day in July. This automatically leads to a higher demand for the sanitary service to accommodate all these patients, and thus to an increase in the number of firemen interventions.

A further step to improve model accuracy is to sort the firemen interventions into normal and abnormal values. After detecting 1122 anomalies and replacing them with the mean of the non-outlier values, better prediction was achieved for the entire dataset and for both the pre and post COVID-19 periods (indicated by a reduction in MAE and RMSE). On the top of this, the most important result is that after applying modification to the dataset, the breakpoint always remains the same. This shows that it is impossible to ignore the period of coronavirus: Even after replacing all anomalies, the COVID-19 period was always detected. Nevertheless, it is a matter of great concern to verify the predictive accuracy

of the interventions classified as abnormal. After replacing the anomalies with the mean of normal values, the results of MAE 0.762 and RMSE 1.995 are very prosperous.

6 Conclusion

The aim of this study is to predict the number of times firefighters are called after detecting a breakpoint associated with the COVID-19 epidemic. The first step was to select features by applying the feature importance method, then the breakpoint was detected and last anomalies were replaced. In each step, the prediction of firemen interventions was performed using the XGboost algorithm. The results show promising accuracy in predicting regular or irregular events such as the COVID-19 epidemic. For future work, prediction of the number of firefighters' mission by intervention type will be implemented as well as other machine learning techniques to optimize the accuracy of this dataset.

Acknowledgement

This work has been supported by the EIPHI Graduate School (contract ANR-17-EURE-0002) and is partially funded with support from the Hubert Curien CEDRE programme n° 46543ZD.

References

1. Cerna, S., Arcolezi, H., Guyeux, C., Royer-Fey, G., Chevallier, C. (2021). Machine learning-based forecasting of firemen ambulances' turnaround time in hospitals, considering the COVID-19 impact. Applied Soft Computing, 107561.
2. Nahuis, S., Guyeux, C., Arcolezi, H., Couturier, R., Royer, G., Lotufo, A. (2019). Long short-term memory for predicting firemen interventions. In 2019 6th International Conference on Control, Decision and Information Technologies (CoDIT) (pp. 1132–1137).
3. Couchot, J.F., Guyeux, C., Royer, G. (2019). Anonymously forecasting the number and nature of firefighting operations. In Proceedings of the 23rd International Database Applications Engineering Symposium (pp. 1–8).
4. Guyeux, C., Nicod, J.M., Varnier, C., Al Masry, Z., Zerhouny, N., Omri, N., Royer, G. (2019). Firemen prediction by using neural networks: A real case study. In Proceedings of SAI Intelligent Systems Conference (pp. 541–552).
5. Elias Mallouhy, R., Guyeux, C., Abou Jaoude, C., Makhoul, A. (2021). Time Series Forecasting for the Number of Firefighters Interventions. In International Conference on Advanced Information Networking and Applications (pp. 39–50).
6. Sujath, R., Chatterjee, J., Hassanien, A. (2020). A machine learning forecasting model for COVID-19 pandemic in India. Stochastic Environmental Research and Risk Assessment, 34, 959–972.
7. Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., Gloaguen, R. (2020). COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. Mathematics, 8(6), 890.

8. Kassania, S., Kassanib, P., Wesolowskic, M., Schneidera, K., Detersa, R. (2021). Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning based approach. Biocybernetics and Biomedical Engineering, 41(3), 867–879.
9. Keshavarzi Arshadi, A., Webb, J., Salem, M., Cruz, E., Calad-Thomson, S., Ghadirian, N., Collins, J., Diez-Cecilia, E., Kelly, B., Goodarzi, H., others (2020). Artificial intelligence for COVID-19 drug discovery and vaccine development. Frontiers in Artificial Intelligence, 3, 65.
10. Zhang, T., Wu, Q., Zhang, Z. (2020). Pangolin homology associated with 2019-nCoV. BioRxiv.
11. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery data mining (pp. 2623–2631).
12. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., others (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1–4.
13. Geurts, P., Ernst, D., Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3–42.
14. Segal, M. (2004). Machine learning benchmarks and random forest regression. Journal is required!.
15. Truong, C., Oudre, L., Vayatis, N. (2020). Selective review of offline change point detection methods. Signal Processing, 167, 107299.
16. Liu, F., Ting, K., Zhou, Z.H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1), 1–39.
17. Ali, M. (2020). PyCaret: An open source, low-code machine learning library in Python. PyCaret version, 2.
18. Zien, A., Krämer, N., Sonnenburg, S., Rätsch, G. (2009). The feature importance ranking measure. In Joint European conference on machine learning and knowledge discovery in databases (pp. 694–709).