Detection of Dangerous Maritime Refugee Migration Ways through Cell Phone Activities

Mustafa ÇOBAN[a,*], Hüseyin Oktay Altun[b], Semih YUMUŞAK[a], Yasin Yılmaz[c]

[a]*Department of Computer Engineering KTO Karatay University, Akabe, Alaaddin Kap Cd. No:130, 42020 Karatay/Konya*
[b]*Department of Electrical-Electronics Engineering KTO Karatay University, Alaaddin Kap Cd., 42020 Karatay, Konya, Turkey*
[c]*Department of Electrical Engineering, University of South Florida,Tampa, FL 33620*

Abstract

During 21st century, world have witnessed devastating wars causing people to migrate, which caused problems in receiving countries. Among those migration routes, maritime migration paths are more preferred compared to other routes since the coasts can not be controlled as strictly as the alternative passages. However, maritime migration poses life-threatening risks due to traveling via unsafe boats, transportation between undesignated areas, lack of emergency life-saving equipment and choice of dangerous weather condition times for covert operation. These migrations refugees and migrants dying or missing at seas. Most refugees are unaware of the high risks they take while they hopelessly choose to move to better life standards. In this study, we propose that, such suicide-like maritime migration activities can be captured to some extent via cell phone activities and there may be a possibility to track early signs of immigrants coming together from other regions and migrate through seaways. By collecting the news from media about unsuccessful attempts of immigrants and relating with the D4R cell phone data provided, we reached some evidence on the possibility of prospective early warning systems through mobile call analysis.

*Corresponding author.
Email addresses: mustafacoban994@gmail.com (Mustafa ÇOBAN), oktay.altun@karatay.edu.tr (Hüseyin Oktay Altun), semih.yumusak@karatay.edu.tr (Semih YUMUŞAK), yasiny@usf.edu (Yasin Yılmaz)

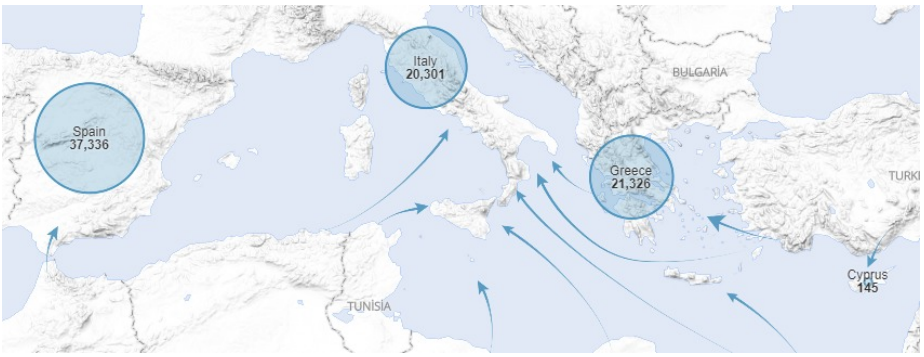


Figure 1: Main maritime immigration routes to Europe. Numbers indicate the tally of sea arrivals at 2018 till mid September, according to UN Refugee Agency. [1]

Keywords: Safety and security of immigrants - D4R - Maritime refugee mobility - Early detection of unsafe migration pathways

1. Introduction

Written human history is full of wars, invasions, violence and migrations. Unfortunately, extraordinary advances in science and technology in the last three centuries of human history did not improve things much, and humanity is still suffering with wars and migrations. Hence, refugees still exist to seek a better future, which is not a crime but rather a human right. Middle East is one of the frequently wounded delicate regions of 21st century, most recent humanitarian crisis existing in Syria for a prolonged time. According to UN more than 5 million people were relocated out of Syria, majority of whom moved to Turkey. Some for family reunification, others for better economic and educational opportunities; a substantial amount of these refugees have been moving to Europe. However, formal ways of being accepted to European countries is highly unlikely due to many official restrictions and quotas.

2. Related Works

Under restrictions and quotas imposed by European countries, most immigrants are taking unofficial and unauthorized attempts in order to proceed

Previous years	Sea arrivals	Dead and missing
2018	74,388	1,642
2017	172,301	3,139
2016	362,753	5,096
2015	1,015,078	3,771
2014	216,054	3,538

Table 1: Number of refugees arrived to their destinations via maritime routes and number of dead and missing people according to UN Refugee Agency. [1]

to Europe. Among these attempts, maritime migration attempts are highly preferable due to lack of control over very long shorelines between countries. One such long coast line is between Turkey and Greece, and thousands of immigrants are moving towards Europe through Greece, Italy and Cyprus. Fig. 1 shows the mainstream maritime immigration routes to Europe according to the UN Refugee Agency. The numbers on the map indicate the total number of sea arrivals since the beginning of 2018 till the midst of September. Compared to the land arrivals being 4720 for the same period, the sea arrivals is almost twenty folds of land arrivals. And unfortunately, for the same period again, 1640 people are reported dead or missing during maritime migration attempts. Out of every 70 maritime immigrants, one of them has lost her life, which is an extraordinarily high risk of mortality in the presence of safe modern transportation vehicles. Table 1 provides number of sea arrivals versus number of dead and missing maritime immigrants between 2014-2018. [1]

In this study, we investigate the links between cell phone activities of refugees in Turkey provided via D4R challenge and potential maritime refugee tracks or timings; in order to detect such events, prevent them making extremely risky voyages via technically insufficient vehicles or at least inform them about the risks they are about to take on behalf of themselves or their loves ones. It must be noted at this point that the twenty five percent of these maritime

refugees are children. For this purpose, we collected data from internet portals of mainstream media or internet media, mostly covering unsuccessful attempts of maritime immigrants being caught by authorities at shores or during the voyage in the middle of the sea. We have proved to some extent that the cell phone data can be utilized in order to allocate the unexpected gatherings of immigrants along shorelines.

3. Methodology

This section provides the methodology on how this study utilizes the Turk Telekom D4R dataset [2] to analyze the refugee communication network. The voice and SMS communication patterns are explored based on unsupervised (Anomaly detection) and supervised (regression classifier) methods. Additionally, a real time integration of the proposed methods is explained as a concept.

3.1. Anomaly Detection

Where as the sms/call durations and counts are an aggregated summary of the raw data, it is yet applicable to be fitted as a time series model based on hours and days. The time series models[3] can be applied in different forms such as auto-regressive or moving-average models to detect anomalies in the time series data. In order to detect an anomaly in a time-series data collection, moving average can be applied as: $s_i = \frac{1}{n} \sum_{j=i}^{i+n-1} a_j$. Fig. 2 is an illustration of a moving average curve, which may create different curves based on the definition of an average function and a window size.

When the moving average model is applied on the D4R voice dataset, the call durations can be visualized and smoothed to see the seasonal change in call durations. As an unsupervised method, the moving average model is applied on the D4R dataset to detect anomaly days/hours for cities, counties, and base stations. The applied model for the D4R data can be accessed in the project repository[4].

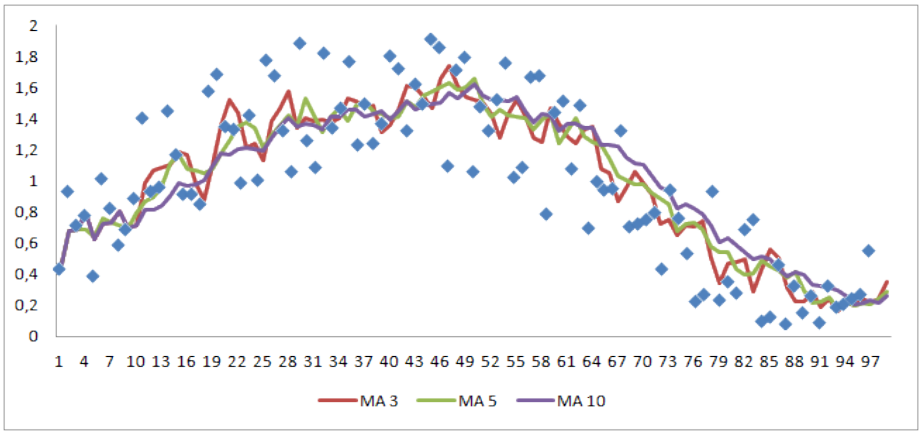


Figure 2: Sample moving average curves for different window sizes

3.2. Regression Classifier

In this study, different type of regression classifiers were experimented. Instead of using only the D4R dataset as the feature set, we have also collected news/events from media sources and created a event-labeled dataset. The details of these datasets are explained more in detail in Section 5. As the numeric classifiers, different regression models were implemented. A regression model is explained as $Y \approx f(X, \beta)$ where the model relates Y and a function of X and β. The regression model can be applied in many forms such as; linear, logistic, multinomial, and polynomial etc. We have experimented the regression classifiers by using the labels assigned by the event extraction process. The event extraction process is explained in section 5. During the classifier verification process, the labeled data is divided into training (70%) and test (30%) sets.

4. D4R Data

The dataset provided by D4R [2] consists of four main segments. Dataset 1 contains the total durations, counts, and base station information about SMS and call transactions. In this study, we have made experiments by using incoming/outgoing communication data in Dataset 1, which are segmented by date

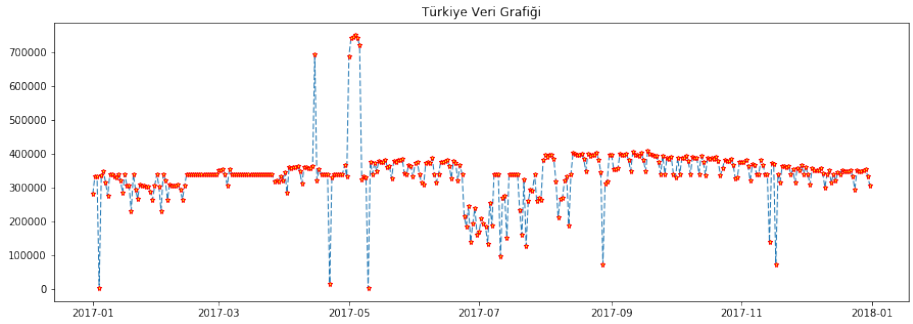


Figure 3: Refugee voice call counts per day for the whole country

and hour. By the help of the refugee labels in Dataset 1, the communication traffic can be monitored.

In Dataset 2, anonymized call records are provided, which contain base station information about individual calls in a 14 days period. Every 14 day period contains randomly generated user ids. By using Dataset 2, the locations of users were identified. Dataset 2 is used to to monitor individuals calling illegal migration areas. Dataset 3 provides location information about the base stations and related county/city information. By using this dataset, we created a geographical map view to determine the refugees living in rural areas.

62 days out of 365 days are missing in Dataset 1 and there are many hourly data missing as well. We extracted the calls/SMS aggregations for base stations, counties, and cities in Turkey. To visualize the characteristics of the data, we created diagrams from the aggregated data. Fig. 3 shows the total call counts of refugees for the whole country, which demonstrates the number of call changes for the whole year and can be replicated for every single location (base station, county, or city).

Additionally, the places where refugees mostly live can be extracted from the D4R data. For instance, by aggregating the call/sms counts, it is possible to identify the living places for refugees. In Fig. 4, the density of refugees in rural areas are visualized.

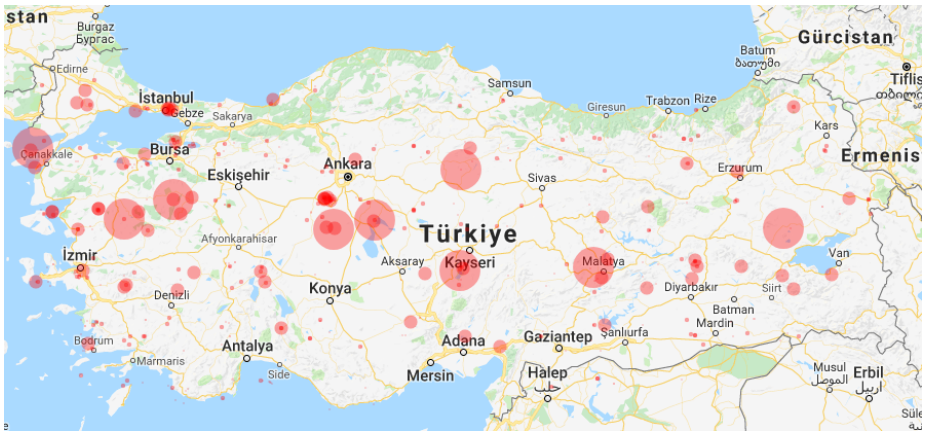


Figure 4: Refugee call densities living in rural areas

By using the routed information contained in the D4R data, it is also possible to create a routed communication graph. By aggregating the total number of calls between base stations, the highest number of refugee voice calls are visualized in Fig. 5.

Figure 5: Top 20 count of voice calls between base stations (weighted connections)

As seen in Fig. 5, whereas the highest number of incoming/outgoing calls are related to Istanbul, the communication around Izmir region, Konya-Adana region, and Van-Gaziantep region indicates the refugee flow and connections in Turkey.

By using these communication data, many experimentation were done on refugee related information. In this study, experiments indicating refugee im-

migration patterns are reported in the following section.

5. Experimental Results

In order to test the hypothesis that refugees fleeing via maritime transport may alter the statistics of cell phone call features of towns located at sea shorelines, we collected data from the internet: websites of local or national media (e.g. Hürriyet API[1]). We used the keywords tabulated at Table 2.

List of search keywords in:		
Mülteci	Kaçış	Kaçma
Yakalama	Kaçış Organizasyonu	Kaçarken Yakalandı
Ege denizi kaçış	Ege'de kaçarken	Kaçma planı

Table 2: Keywords to search refugee maritime flee attempts

The search results have been examined and the related news is skimmed to a table indicating if some group of refugees have been detected at some specific date and location, or not. Table 3 tabulates a sample of collected data. Based on the search results of the news agencies, we have determined some hot-spot towns such as Bodrum, Ayvacık, Didim, Çeşme and Reyhanlı etc. We then get the corresponding data from D4R database and graphed the voice call count according to the days of the year. The number of voice calls at Bodrum throughout the year involving at least one refugee is illustrated at Fig. 7.

Based on the events/migrations list, a linear regression classifier is applied into the labeled SMS/voice call records. In this section, we provide the experimental results for a specific location, which is considered to be the main route of illegal immigration. Other results for other routes are also available in the project repository [4].

The changes in the refugee-to-refugee call counts were analyzed by applying the regression classifier to the labeled dataset. Before applying the classifier on

[1] https://developers.hurriyet.com.tr/

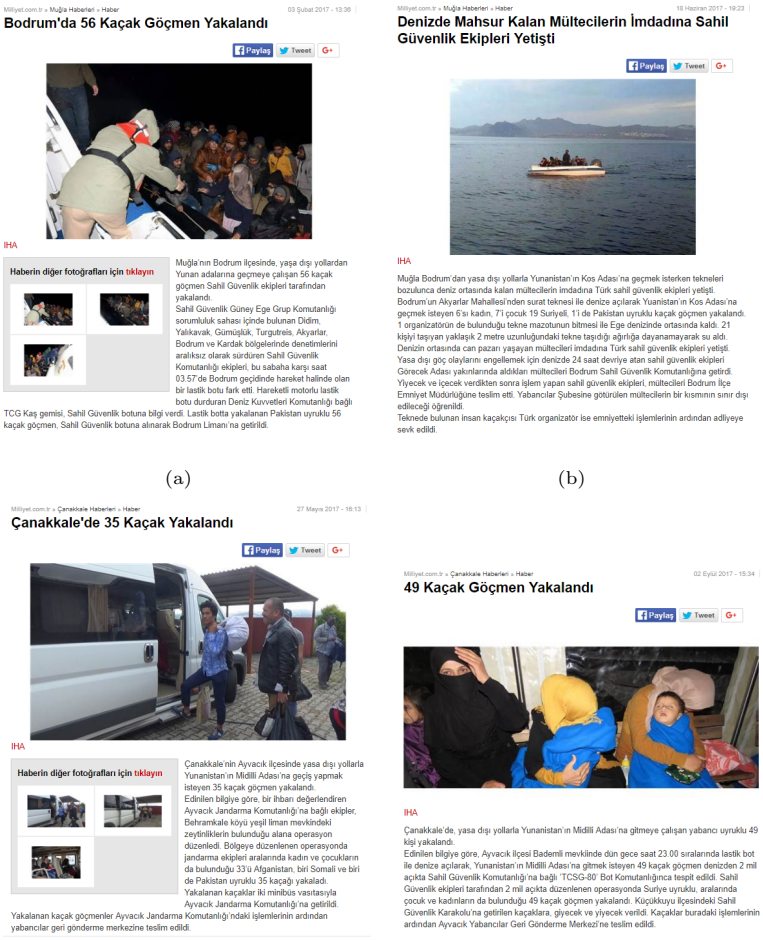


Figure 6: Some search results depicted from the internet. All the sources are documented at our project's GitHub site. [4] News are collected from Milliyet web site. [5]

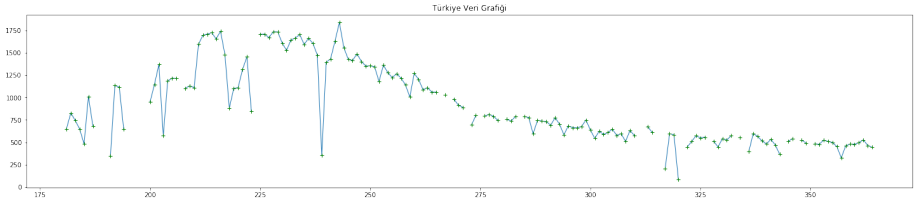


Figure 7: Data related to Bodrum: Voice Count vs. days of the year

Date	City	District	event
2017-01-03	Muğla	Bodrum	1
2017-01-02	Çanakkale	Ayvacık	1
2017-11-17	Aydın	Didim	1
2017-05-22	İzmir	Çeşme	1
2017-05-13	Aydın	Kuşadası	1
2017-09-15	Balıkesir	Ayvalık	1
2017-01-09	Hatay	Reyhanlı	1
2017-08-25	İzmir	Dikili	1

Table 3: Data is collected and organized as in this table. If there exist a news about an unsuccessful flee attempt at Turkish side, we indicate the date and location as one, and zero otherwise.

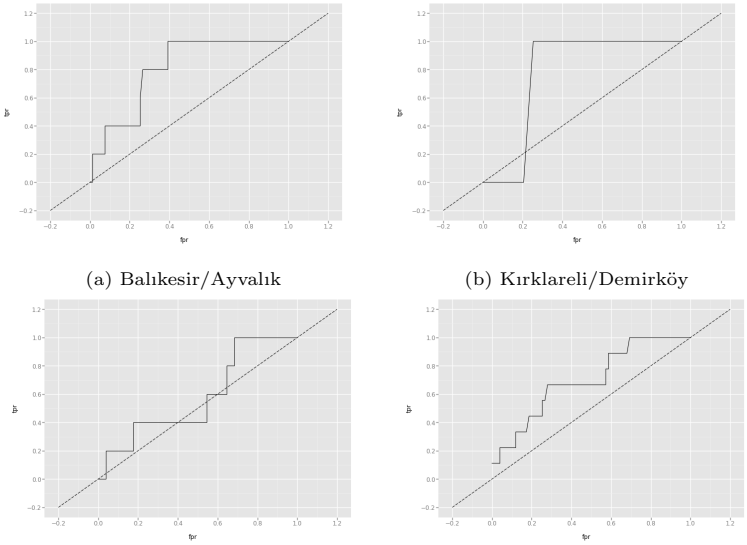
voice and SMS counts, a missing data imputation (padding) was also performed for the missing dates. The ROC results for voice communication without a missing data imputation are visualized in Fig. 6, with missing data imputation in Fig. 9 Additionally, for the SMS message traffic, with missing data imputation the ROC curves are visualized in Fig. 10 and without imputation in Fig. 11.

As experimented in this section, during an illegal immigration activity, there exists a correlation with number of the SMS and voice calls. Therefore, it is proposed that the change in the SMS and voice calls can be used as an early alert method to detect a near-future illegal immigration.

5.1. Analysis of the Base Stations on the Immigration Routes

To deeply analyze the issue for specific locations, single base stations are analyzed to detect early signs of an illegal immigrant activity. For this reason, base stations near one of the main routes (Bodrum/MUGLA) are analyzed.

The voice count data for Bodrum is visualized in Fig. 7, which includes several up and downs in the graph. As extracted from the base station table, there are 30 different base stations located in Bodrum. All Bodrum related



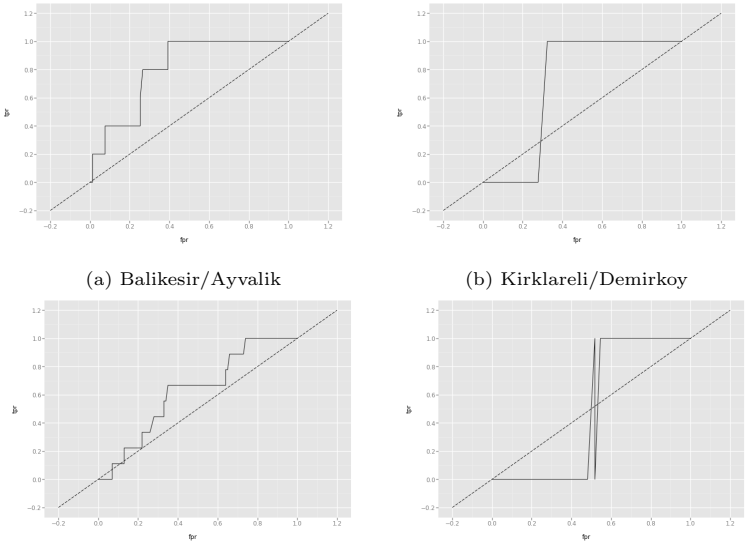
(a) Balıkesir/Ayvalık

(b) Kırklareli/Demirköy

(c) Aydın/Kuşadası

(d) Muğla/Bodrum

Figure 8: ROC of which voice data with illegal immigration dates



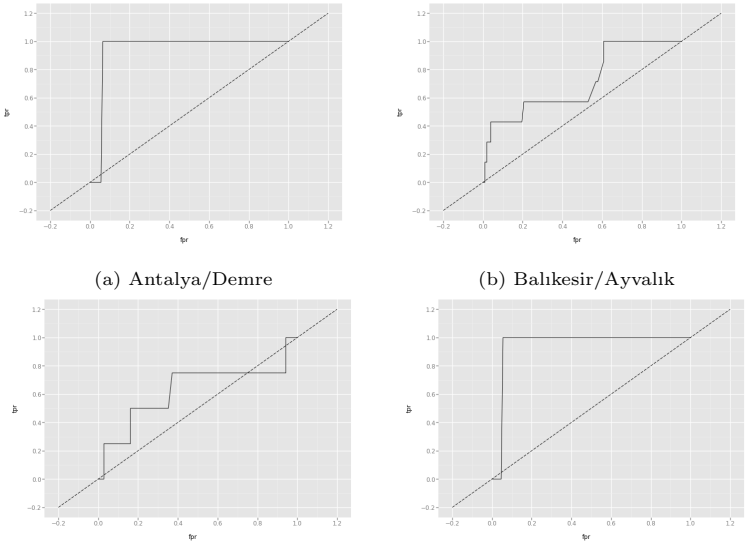
(a) Balikesir/Ayvalik

(b) Kirklareli/Demirkoy

(c) Muğla/Bodrum

(d) Mugla/Marmaris

Figure 9: ROC of which fill voice data with illegal immigration dates



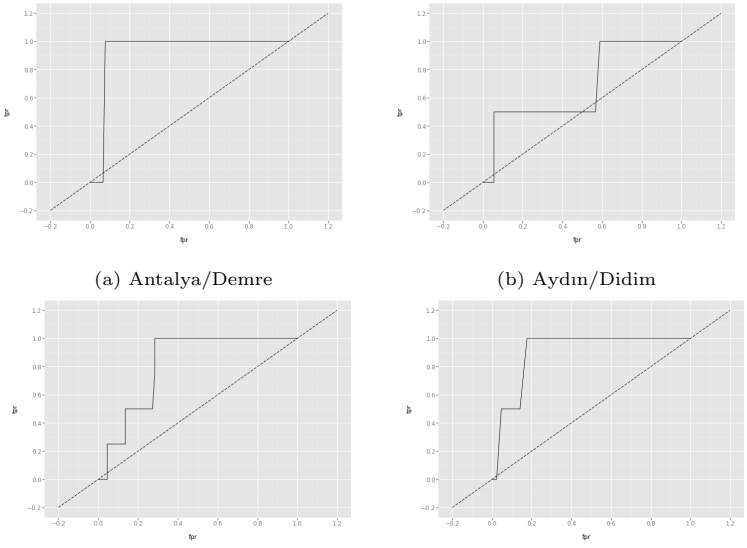
(a) Antalya/Demre

(b) Balıkesir/Ayvalık

(c) İzmir/Dikili

(d) Muğla/Marmaris

Figure 10: ROC of which fill SMS data with illegal immigration dates



(a) Antalya/Demre

(b) Aydın/Didim

(c) Aydın/Kuşadası

(d) Kırklareli/Demirköy

Figure 11: ROC of SMS data with illegal immigration dates

immigration events are analyzed on these base stations and it is identified that three base stations have a correlation with the events. Some of the results for the classification results are visualized as an ROC curve in Fig. 12

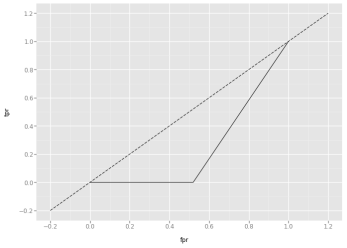
With the help of the geographical coordinates provided in the D4R dataset, the illegal immigration correlated base stations are shown in Fig. 13

6. Discussions and Conclusions

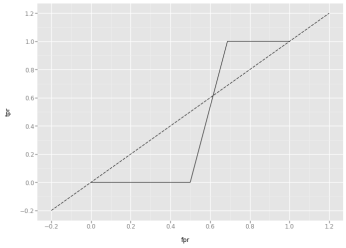
In this study, we have gathered evidence from the provided D4R data that the unsuccessful attempts of maritime travel have some correlation at Turkey's certain shore towns' refugee mobile phone activities. The results suggest certain degree of correlation, however correlation is not very strong. We can argue two possible reasons, one fact being substantial amount of D4R data is missing. The other factor is we only collected data from Turkish news sources for unsuccessful attempts, however a symmetrical approach could have been performed in order to gather data from International news resources in order to account for successful events. We believe that improving either of these conditions may capture a stronger correlation to support our hypothesis for future works.

When we collected generous amount of data for the presence of unsuccessful attempts to flee from a shore town via news resources and the ROC curves ended up being weak showing no correlation; we zoomed further into the data of cell towers in that town, and realized that some towers are showing correlation and nicer ROC characteristics. We reason that, for some of the shore towns, refugee population is not scattered across the town but concentrated rather in a small region. A more detailed data collection can reveal us some popular immigrant cafes and social places, so that we can proceed even further level of data analysis.

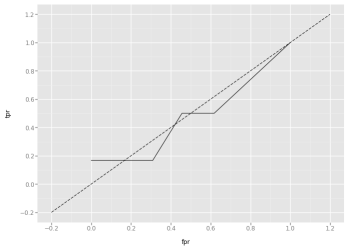
Here we want to discuss a bit on integration concept of prediction methods in real time systems. Although the predictive and prescriptive methods performs well in offline databases, it is also crucial to have a real time integration with the GSM infrastructure. In order to simulate this integration, we created a concept to integrate our algorithms with the D4R infrastructure by the help of an AI



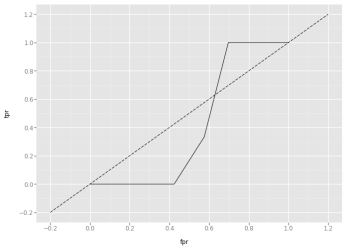
(a) Base station: 5064923



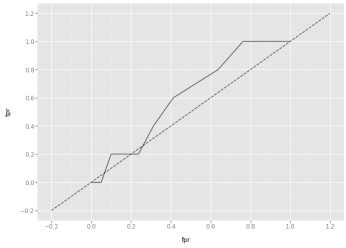
(b) Base station: 5064978



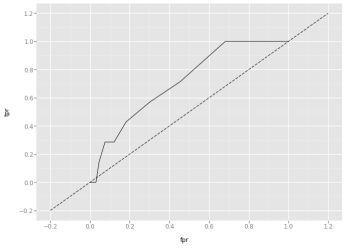
(c) Base station: 5066687



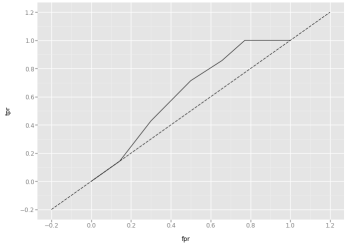
(d) Base station: 5067295



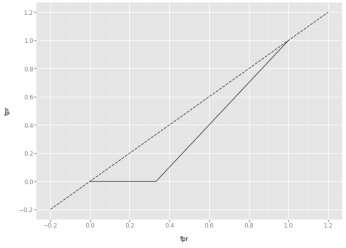
(e) Base station: 5165506



(f) Base station: 5067296



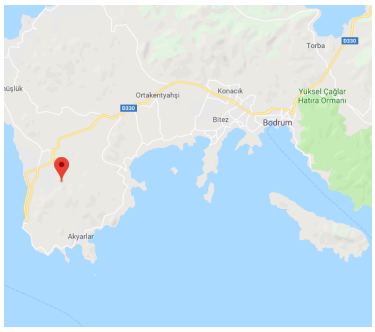
(g) Base station: 5066686



(h) Base station: 5064984

Figure 12: ROC for some base stations in Bodrum

(a) Location: 5066686



(b) Location: 5067296

(c) Location: 5165506

Figure 13: Locations of three base stations correlated with Refugee migrations

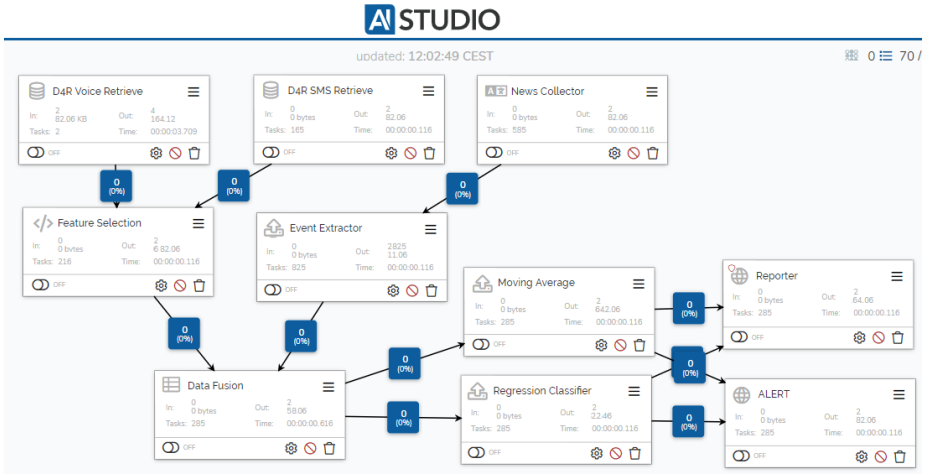


Figure 14: AI studio integration

Studio Toolbox[2]. In Fig. 14, the designed data flow of our integration patterns and AI processing is illustrated. The data processing flow in Fig. 14 may perform the following steps; a)Data retrieval: Currently from the SQL databases created for D4R, b)feature selection, cleansing, and missing data amputation, c) news collection and event extraction, d) data labelling and fusion, e)analytics: anomaly detection and regression classifier and f) alert and reporting.

Here we presented a novel approach for maritime travel detection and obtained some promising results. If used correctly, this technology may prevent hundreds of refugees being drowned, or at least make more informed and rational choices when they are at the verge of making a terrible decision. We believe that seeking refuge from suppression or violence is a basic human right and should be used to the full extent when necessary, and have no intentions to propose these techniques to prevent immigrants to pursue happiness and freedom.

7. Acknowledgements

We would like to thank Dr. Martin Voigt for his support on using the AI Studio toolbox and valuable comments, Dr. Pin-Yu Chen and Dr. Sijia Liu for the ideas about machine learning algorithms, Dervise Nur Gokalp on her commitment for creating the event news dataset and Mehmed Bugrahan Duran, Burhan Toruk, Sevket Ozkaya, for their technical support.

References

[1] UN Refugee Agency, Operational Portal, Data2.unhcr.org. (2018). Situation Mediterranean Situation. [online] Available at: `https://data2.unhcr.org/en/situations/mediterranean` [Accessed 15 Sep. 2018].

[2]http://www.ai4bd.com

[2] Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dağdelen, Ö., 2018. Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523.

[3] Durbin, J. (1960). The Fitting of Time-Series Models. Revue De L'Institut International De Statistique / Review of the International Statistical Institute, 28(3), 233-244. doi:10.2307/1401322

[4] D4R Challenge Solution Repository. (2018). [online] Available at: `https://github.com/semihyumusak/D4R_Challenge_Solution` [Accessed 15 Sep. 2018].

[5] Milliyet Haber - Türkiye'nin haber sitesi. (2018). Milliyet Haber - Haberler. [online] Available at: `http://www.milliyet.com.tr/` [Accessed 16 Sep. 2018].