Mobile Internet Activity Estimation and Analysis at High Granularity: SVR Model Approach

A. Rizwan*; K. Arshad †; F. Fioranelli*; A.Imran ‡; M. A. Imran*
*School of Engineering, University of Glasgow, Glasgow UK
†Faculty of Engineering, Ajman University, UAE
‡Gallogly College of Engineering, University of Oklahoma, Tulsa, USA
Email: a.rizwan.1@research.gla.ac.uk; k.arshad@ajman.ac.ae †; ali.imran@ou.edu ‡;{francesco.fioranelli; muhammad.imran}@glasgow.ac.uk*

Abstract— **Understanding of mobile internet traffic patterns and capacity to estimate future traffic, particularly at high spatiotemporal granularity, is crucial for proactive decision making in emerging and future cognizant cellular networks enabled with self-organizing features. It becomes even more important in the world of 'Internet of Things' with machines communicating locally. In this paper, internet activity data from a mobile network operator Call Detail Records (CDRs) is analysed at high granularity to study the spatiotemporal variance and traffic patterns. To estimate future traffic at high granularity, a Support Vector Regression (SVR) based traffic model is trained and evaluated for the prediction of maximum, minimum and average internet traffic in the next hour based on the actual traffic in the last hour. Performance of the model is compared with that of the State-of-the-Art (SOTA) deep learning models recently proposed in the literature for the same data, same granularity, and same predicates. It is concluded that this SVR model outperforms the SOTA deep and non-deep learning methods used in the literature.**

Keywords— Big Data Analytics, Mobile Internet Traffic Estimation, High Granularity Spatiotemporal Analysis, SVR

I. Introduction

During the last decade, mobile services have sharply evolved from only cellular network-based services like messages and calls to internet-based services like mobile apps and web surfing on mobiles. On one end such services demand a different set of bandwidths, network protocols and resources for the transmission of diversified data types, there on the other end, they have also raised user-network interaction to the highest level ever and this trend is increasing [1]. Data consumption, user interaction with the network and the time spent by users on the cellular network to access Internet-based services, has surpassed the conventional cellular services such as call and Short Messaging Service (SMS). StatCounter, a research company that tracks internet activity globally, concluded that the number of web pages accessed using mobile devices already exceeds the number of web pages accessed from Personal Computers (PC) and laptops in October 2016, and this trend is also increasing [2]. For instance, subscribers from the USA spent almost 90% of their mobile phone time on the mobile internet in 2015.

These statistics clearly show an increasing demand of a huge range of Internet-based services on the cellular network and require the network to be capable to cater a variety of data types with efficiency and better latency [3]. So, it becomes crucial for the network to learn user's internet usage behavior and preferences in terms of contents, timings, and vicinity for the provision of user-specific seamless services. Further to it, the future network must be able to predict demand for internet services at different spatiotemporal granularity for better Radio Resources Management (RRM) and pre-emptive measures against key challenges like admission control, traffic congestion etc. To meet that objective, there exists a need to design most efficient and optimal RRM algorithms. The analysis of spatio-temporal patterns of internet consumption at higher granularity is also important for the understanding and information management of varying communication level expected among numerous devices locally in future networks [4].

The increased frequency of user network interaction has also led to an activity level of very high granularity over the network with fine footprints of respective activity records, e.g CDRs. Such CDRs, also provide an opportunity to gather intelligence about users' behaviors and preferences towards different on and off network services. This intelligence can be accumulated by the identification of patterns and correlations in the existing data with the application of data analytics. Such data analytics based cognizance can help to improve overall network performance via RRM strategies at shorter intervals by making timely autonomous decisions [5]. This kind of network intelligence is a driving factor to make future networks

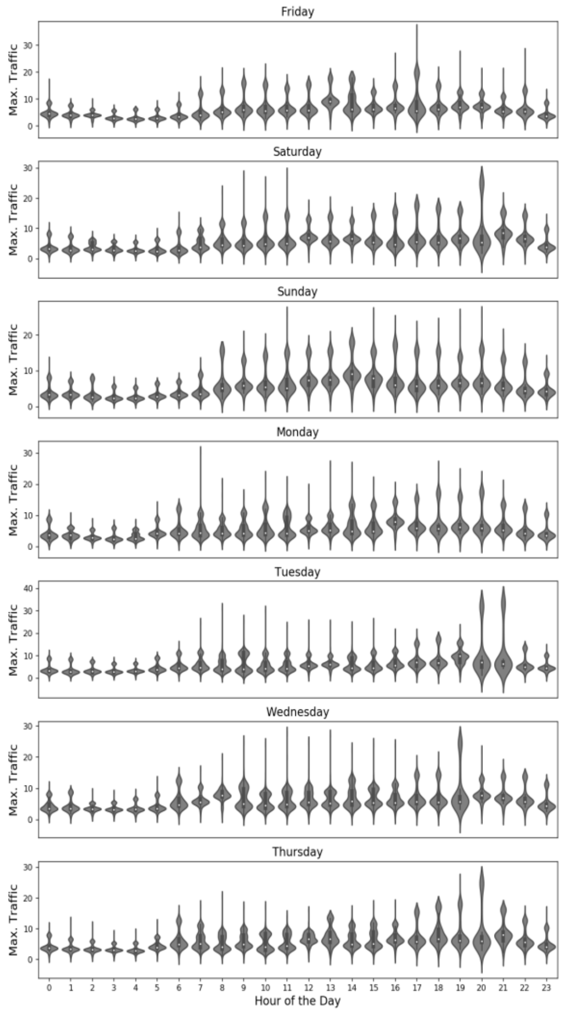

Fig 1: Cells distribution with respect to maximum traffic generated on hourly basis in a week

more pre-emptive, autonomous and self-organizing, some of the key features expected in the future cellular (i.e. 5G and beyond) and IoT based networks [6].

Spatiotemporal understanding and prediction of traffic can help optimize resources like switching off certain eNodeB for possible energy conservation. Similarly, timely and accurate traffic prediction can also play an important role in managing operational and quality of services related problems e.g. congestion control, admission control, network bandwidth allocations etc. [7]. Previously we proposed a small cell sleep cycles centered approach [8] that leverages from spatiotemporal prediction based on same CDR to pro-actively schedule radio resources. Results for this approach show substantial energy savings and reduced inter-cell interference (ICI), without compromising the users Quality of Service (QoS). Besides future traffic prediction, understanding of high granularity spatiotemporal traffic patterns and the distributions are also important for network planning and configuration

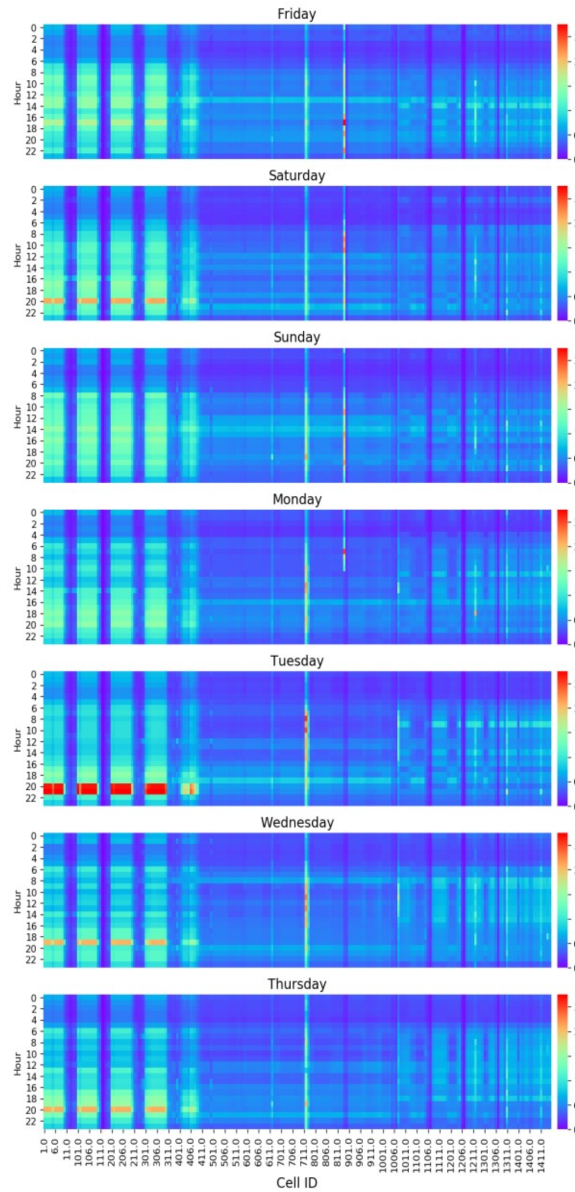

Fig 2: Heat map of maximum hourly internet activity in 225 cells over a week

for future networks where network densification is seen as a mean to meet the diversified high data demands.

In this paper, the mobile users' internet usage behavior with respect to time and location is studied with the help of real network data. For the purpose of analysis, actual two months cellular internet activity data for Milan city, released by Telecom Italia, is used. This paper explores the real internet traffic variance over a network in the spatiotemporal domain at a high granularity level particularly in term of time where the the variance of traffic even within an hour is studied. The detailed study of the users' preferences in terms of the data contents is out of the scope of this paper.

Cellular data is a rich source of information for multidisciplinary research and multifaceted decision-

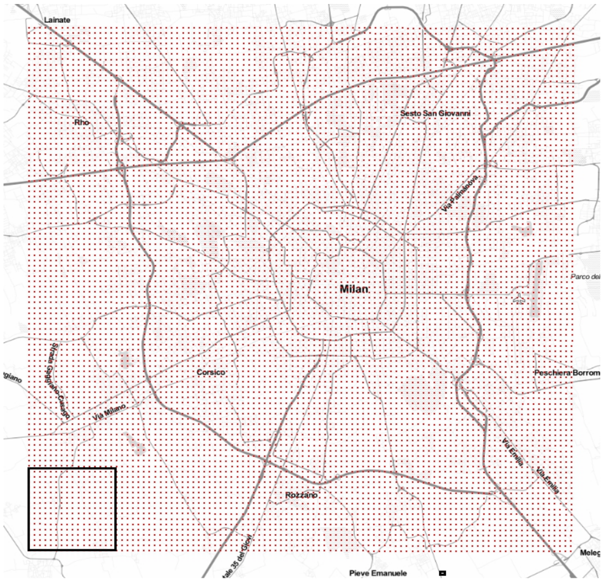

Fig 3: Grid over Milan and area under observation

making processes. There exists enormous research on cellular network architecture, functionalities, and services. Now plenty of work also exists on the utilization of data analytics for network improvement as it can be seen from [8], [9] and [10], where [10] specifically discuss predictability in networks. Significance of the use of real mobile data in analytics is very high as it captures and exhibits the true feel of actual network behavior. But research on spatiotemporal analysis and predictability of cellular network traffic based on real network data is very limited and research on spatiotemporal analysis and predictability at high granularity is even rare especially for the internet traffic. Authors in [1] emphasize the need of models to predict traffic demand on short (e.g., minute to hours) and medium intervals of time (e.g., days to weeks) after presenting a detailed review of the literature published in the last decade on the topic.

Another drawback of the existing research is that aggregated hourly activity level is taken into account whereas maximum traffic is more significant for estimation of demand and resource allocation [11]-[13]. The aggregated or mean traffic is more stable as compared to maximum traffic which has high spatio-temporal variance as it can be seen from Fig.1 and Fig.2. Further, most of the traffic prediction models proposed in the literature are for call and SMS data only and were separately trained and tested at different locations [11]-[13]. In this research, our proposed model is platform and location independent. The model is trained simply by providing six data points of aggregated activity level, each for ten minutes in an hour, for all cells.

In the literature, Artificial Neural Networks (ANN) is one of the most popular non-linear models to forecast complex network traffic and outperform traditional time-series models like ARMA and FARIMA [14]. Studies focusing internet traffic on the cellular network for spatiotemporal analysis and short-term predictability are very rare. In [15], authors have applied deep learning methods for the prediction of internet traffic and results are used as a benchmark in this paper.

In this paper, a Support Vector Regression model is used for the prediction of future internet activity for three different levels, minimum, maximum and mean at high granularity. These levels help to have a basic idea about the activity level in the different cells for a shorter period of time. The performance of the proposed method is compared with the SOTA deep learning methods available in the literature. We aim to prove that a classical, comparatively simple, SVR model can perform much better than the complex deep learning models for cellular network problems like the one under study here, internet activity estimation at high granularity.

Deep learning models are not the optimal solution in all cases, therefore we focus on these three predicate tasks (activity levels) used by authors in [15] for their deep learning models and compare the performance of SVR with that of the deep learning models for the same data, granularity, predicates and evaluation metrics. This is a timely research as future internet activity estimation using data mining and machine learning over a cellular network at high granularity is one of the most important problems for the research community in order to design efficient and intelligent 5G and beyond 5G cellular networks [16]. In this paper, it is concluded that the proposed SVR based method outperforms SOTA approaches used in the recent literature.

The remainder of this paper is organized as follows: Section II describes the dataset used in our analysis. Section III explains the methodology used for data analytics and model training, testing and performance evaluation. Section IV introduced the proposed SVR model and the performance metrics used in this paper and describes how the model is implemented. Section V presents the analysis and results of our proposed model. Finally, section VI concludes this paper.

II. DATA SET DESCRIPTION

To study the internet activity dynamics on a cellular network it is of paramount importance to use actual data from a cellular network operator. The internet activity data used in this paper is obtained from a comprehensive big dataset released by Telecom Italia as part of Big Data Challenge 2013 [17]. The dataset includes CDRs (i.e. SMS, call, and internet activity), precipitation data, electricity consumption data, weather station data and website data for the city of Milan, Italy for November and December 2013. In this

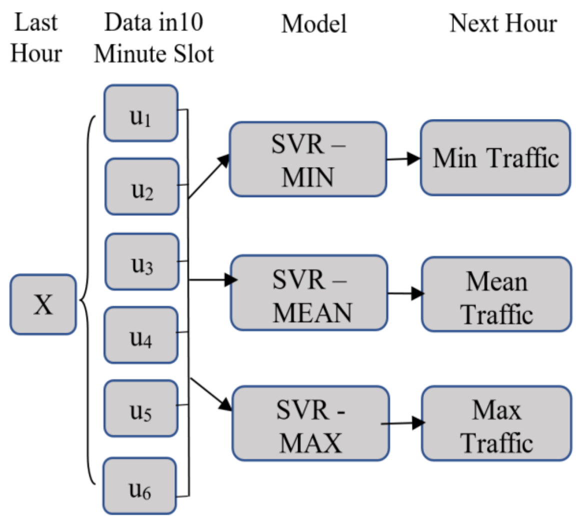

Fig 4: Layout of SVR model implementation

paper, internet activity data from CDRs is used for the spatiotemporal analysis of the behavior of users using smartphones.

For the data collection and aggregation, the city of Milan is geographically mapped as a 100 by 100 grid of 10,000 rectangular cells as shown in Fig.3. Internet activity level is represented by an imitated rational number for confidentiality. Each number refers to an activity level aggregated for each cell separately for a time interval of 10 minutes. These numbers do not represent actual internet data consumption but refer to activity level in a cell. So they can help for the comparative study of internet activity in various cells at different time slots. They can give an idea which cells or time slot have more or less activity as compared to other cells or time slots and how much the difference is. In our experiment, we have used data of nine weeks, for the months of November and December 2013, and 225 cells covering the left bottom corner of the city as highlighted in Fig.3, a grid of 15 by 15 cells. First three weeks data is used for cross-validation and later more data is added for training purposes. Finally, data for the first eight weeks is used for the training purposes and the ninth week data is used for testing and performance evaluation.

III. Methodology

First, we calculated three basic levels of internet activity for each hour of the day for all 225 cells i.e. Minimum, Average and Maximum level of activity. This was followed by the study of the spatiotemporal changes in maximum internet activity level. A density candle plot (Fig. 1) and a heat map (Fig.2) are plotted for the maximum internet activity level which helps to understand maximum internet activity distribution and variance over each day of a week for 225 cells and over a day across the 225 cells respectively. For visualization purposes here focus has been on maximum internet activity as it is commonly the most sought-after feature for network resources planning and allocation.

Three SVR models, independent from each other, are trained and tested for three activities level Maximum, Mean, and Minimum. Initially, SVR models are separately trained using two weeks data and trained models are validated using the data of the third week which was kept separate. In the validation step, SVR exhibited an accuracy of 85%, 87% and 83% for Minimum, Mean and Maximum tasks respectively.

In the end, the model was trained on eight weeks data and tested against unseen data of week nine which was kept separate at the very start of the experiment. The performance of the models was evaluated using the performance metrics defined in section IV.B. The internet activity over an hour, aggregated in six slots of ten minutes each, works as an input to the SVR model for the prediction of each internet activity level for the next hour as shown in Fig. 4.

IV. Proposed Model and Performance Metrics

A. Support Vector Regression (SVR)

We have implemented Statistical Learning Theory based epsilon-insensitive nonlinear SVM regression here. The basic goal of SVR is Structural Risk Minimization (SRM). In practice, the time series of base stations' traffic show non-linear behavior. Hence, non-linear SVR is used here, in the internet activity forecasting scheme.

To formulate the problem, let's define training data as $\{x_i, y_i\}, i = 1,2,3, \ldots, n = 24$ where x_i is the input vector representing an hour of the day (e.g. X_1 represents the first hour in the morning 00:00 to 1:00 am) comprising six scalar values u each representing internet activity for a ten-minute time slot as shown in Fig 4. Similarly, y_i represents the maximum, minimum or mean value of the corresponding hour, depending on task the model is trained for. First, the input is mapped on a multidimensional nonlinear feature space using a non-linear transformation function [18] represented as $\varphi\,(.)$. In such case where we have high dimensional data, regression function can be expressed as follows:

$$f(x) = \omega \,.\, \varphi(x) + b \tag{1}$$

Such that $\omega \in R^d$ and $b \in R$ where d represents the dimensions or number of columns in data as it is 6 in our case and b represents the bias. And outcome of $\varphi\,(x)$ represents the input features space.

The quality of estimation is measured by the loss function. In this paper, the epsilon insensitive loss

TABLE I RESULTS: PERFORMANCE OF MODELS AGAINST METRICS

Task	Metric	SVR	ARIMA [15]	LM [15]	CNN-RNN[15]
Min.	MA	90.25%	67%	61%	69%
	MAE	0.42	22.85	32.86	21.35
	RMSE	0.61	58.18	90.96	50.4
Mean	MA	91%	75%	68%	72%
	MAE	0.45	24.34	33.08	26.85
	RMSE	0.65	52.25	81.46	58.36
Max.	MA	89.72%	63%	63%	67%
	MAE	0.60	49.78	56.01	44.74
	RMSE	0.88	100.85	126.36	92.32

function is used, which ignores errors that are within epsilon (Ɛ) distance of the observed values. For training samples outside epsilon insensitive zone, the slack variable ξ_i, ξ_i^* is introduced that allows the errors to exist up to ξ_i, ξ_i^* beyond epsilon insensitive zone. So SVR model is trained by solving the minimisation problem defined as (2):

$$\min \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$
$$\text{s.t} \begin{cases} y_i - (\omega . \varphi(x_i) + b) \le Ɛ + \xi_i \\ -y_i + \omega . \varphi(x_i) + b \le Ɛ + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \\ i = 1,2,3, \dots, n \end{cases} \quad (2)$$

The constant C is a positive numeric value that regularise the function for flatness and over fitting. It imposes a penalty on the values beyond Ɛ-insensitive zone and determine the level of tolerance for deviation of values beyond Ɛ-insensitive zone. We have used the heuristic method in this paper for the selection of C and Ɛ. By extensive iterations, using the values of C = 1 and Ɛ = .02, the loss function is minimum.

B. Performance Metrics

In order to evaluate the performance of our proposed SVR model, the following performance metrics were used: Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and Mean Accuracy (MA) [15]. The MAE, RMSE, and MA were calculated for each task, Minimum, Mean and Maximum separately. Let y_i represents the actual hourly minimum, mean and maximum internet activity in test data and $\hat{y}_i$ represents corresponding minimum, mean or maximum hourly internet activity predicted by the relevant model. Hence, performance metrics can be written as (3)-(6):

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

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (|y_i - \hat{y}_i|)^2} \quad (4)$$

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

where n represents the number of instances in the test data. Mean accuracy (MA) is measured using MAPE metric as follow:

$$MA = (1 - MAPE) \times 100\% \quad (6)$$

where y_i and $\hat{y}_i$ respectively represent actual and estimated values of minimum, mean and maximum hourly internet activity.

V. ANALYSIS AND RESULTS

As commonly maximum activity level is considered for resources planning and allocation for the networks, therefore for visualization we have displayed plots of maximum internet activity. But we have trained, tested and compared the performance of models for three activity levels.

From cumulative frequency distribution of minimum, average and maximum activity as shown in Fig. 5 it can be seen that it follows power law i.e. most of the cells generate low level activity only a few cells generate high activity, even in CDF for maximum activity level more than 80 percent of the cells have maximum internet activity level less than 10 Same can be inferred from Fig.1 density candles of cells distribution according to maximum activity level over the day for the whole week. It can be observed that for the whole week as a common factor most of the cells have maximum internet activity level around 5. Careful observation of Heatmap in Fig.2 also validates this concept.

Variance in hourly maximum internet activity over the time scale for one week can be observed from Fig.1 where maximum internet activity, on y-axis, is plotted against the hours of the day on x-axis for a week width of the density candles show that at overall in the whole week maximum activity get slower (i.e. less number of cells with much activity) after 11 pm and it remains low till 6 am comparatively. The width of the candles represents the density of no. of cells with maximum activity in that area of the candle with a maximum activity shown on the y-axis for cells in the corresponding area of the candle.

The significant rise can be seen from 7 am, and many of the cells have higher maximum activity level between 8 am and 9 pm as compared to the rest of the day. It can also be seen that there exists a significant second candle on top of the underneath candles for weekdays that reflects some cells have even higher maximum activity on working days as compared to weekends. First candles reflect that most of the cells have maximum activity level value around 5 and some cells have activity level value near 10 on weekdays. On Sunday most of the active cells have internet activity

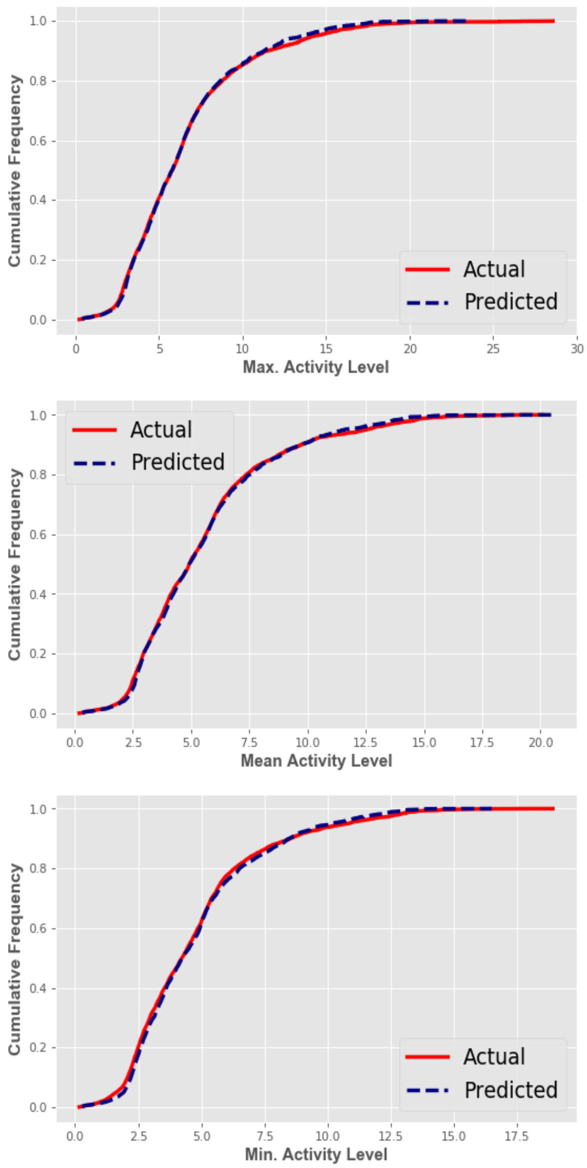

Fig 5: CDF of Mean, Minimum and Maximum Hourly Internet Activity: Actual VS Predicted

level near 5 and width of candles show there are more cells with that much activity compared to other days but second top candles representing cells with even higher activity is not prominent.

Heat map of the whole week in Fig 2. shows a variance of maximum internet activity within the cells represented on the x-axis against each day of the whole week on the y-axis. Poorly active cells are blue with maximum internet activity level between 0 and 6. Cells with maximum internet activity in the range of 6 and 12 falls in blueish-green shade. Substantially active cells with maximum internet activity in the range of 12 and 18 are coloured green. Highly active cells are coloured yellowish having maximum internet activity around 24. Extremely active cells with maximum internet activity level 30 or above represented with red colour. Here again it can be seen that there are few cells which are extremely active in particular time slots, most of the cells do not generate much traffic even at their peak level. Bottom left corner of the selected area generates the highest internet activity. Cells in the range of 1 to 11 and overlying cells 101-111, similarly 201-211, 301-311 and 401-411 which are also consecutive cells generate the highest activity. As we move away from this area internet activity fades out. Beside that heat map also show that most of the cells have higher activity level between 8 am and 9 pm.

While discussing the results of our model SVR, its performance is compared with the performance of deep learning models applied to the same data at similar granularity for similar predicates as in [15]. Authors in [15] also compared the performance of the deep learning models like recurrent neural network (RNN), three-dimensional convolutional neural network (3D CNN) and the combination of CNN and RNN (CNN-RNN) with non-deep learning methods of ARIMA and Levenberg-Marquardt (LM) algorithm based neural networks (NN). In [15] authors recommend Multitask learning CNN-RNN (i.e training and predicting for minimum, mean, and maximum levels in single model) one as a most reliable model that outperforms other deep learning models with the predictability of all levels of internet activity with 70% to 80% accuracy 7% more than that they achieved for single task learning with deep learning methods.

From the comparison of the performance against metrics mentioned above as shown in table 1 it is found that SVR trained an tested separately for all three tasks performs better than all the models proposed by the authors in [15] against all metrics. SVR has an accuracy of 90.25%, 91% and 89.72% for the prediction of minimum, mean, and maximum internet activity respectively which is higher than that of deep learning methods and subsequently from another classical time series model like ARIMA. Root mean square error and mean absolute error are also less compared to deep learning methods for all activity levels. CDF plots in Fig.5 show that estimated and actual minimum, mean and maximum activity have very similar cumulative distribution frequency. That means the same percentage of cells generating similar activity levels. It also can be seen that for minimum hourly internet activity level 60% of the time activity level is less than 5 and in less than 10% instances it is above 10 with an upper bound of approximately 17. For maximum activity tracing 40% of the time activity level is less than 5 and almost 20% of the time it is above 10, it approaches 25 at highest. For mean activity approximately 50% of the times it is less than 5 and approximately 10% of the times it is above 10.

VI. CONCLUSION AND FUTURE WORKS

The capability of understanding and predicting high variance mobile internet activity at a high granularity level is a requirement for autonomous cognizant future cellular networks enabled with self-organizing features. Efficient algorithms are desired to make future cellular networks equipped with this capacity. In our research, we have found that for internet activity over cellular network significant variance is seen in maximum activity over the temporal as well as spatial scale compared to mean or aggregate activity. We implemented statistical learning theory based nonlinear support vector regression model on real network internet activity data at high granularity to predict the maximum, minimum and mean internet activity for the next hour on the basis of internet activity in last hour. We compared the performance of our model with that of the state of the art deep learning and classical models and proved that SVR outperforms the other models.

The results show that SVR algorithm predictions can be used in further research to pre-emptively address practical network problems like traffic congestion. The analysis results here also provide grounds for further research for grouping of cells with similar activity patterns at different granularity and allocation of resources accordingly.

ACKNOWLEDGEMENT

This research is funded by Deanship of Graduate Studies and Research, University of Ajman, project-2017-A-EN-03 and partly by EPSRC Global Challenges Research Fund – the DARE project – EP/P028764/1

REFERENCES

[1] Naboulsi, D.; Fiore, M.; Ribot, S. & Stanica, R. Large-scale mobile traffic analysis: a survey IEEE Communications Surveys & Tutorials, IEEE, 2016, 18, 124-161

[2] gs.statcounter.com/platform-market-share/desktop-mobile-tablet

[3] Gutierrez-Estevez, D.; Bulakci, Ö.; Ericson, M.; Prasad, A.; Pateromichelakis, E.; Belschner, J.; Arnold, P. & Calochira, G. RAN enablers for 5G radio resource management Standards for Communications and Networking (CSCN), 2017 IEEE Conference on, 2017, 1-6

[4] Tehrani, M. N.; Uysal, M. & Yanikomeroglu, H. Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions IEEE Communications Magazine, IEEE, 2014,86-92

[5] Romero, J. P.; Sallent, O.; Agusti, R. & Diaz-Guerra, M. A. Radio resource management strategies in UMTS John Wiley & Sons, 2005, pp-177-289

[6] Imran, A.; Zoha, A. & Abu-Dayya, A. Challenges in 5G: how to empower SON with big data for enabling 5G IEEE Network, IEEE, 2014, 28, 27-33

[7] M. F. Iqbal and L. K. John, "Power and performance analysis of network traffic prediction techniques," *2012 IEEE International Symposium on Performance Analysis of Systems & Software*, New Brunswick, NJ, 2012, pp. 112-113

[8] Zoha, A.; Saeed, A.; Farooq, H.; Rizwan, A.; Imran, A. & Imran, M. A. Leveraging Intelligence from Network CDR Data for Interference aware Energy Consumption Minimization IEEE Transactions on Mobile Computing, IEEE, 2017

[9] Blondel, V. D.; Decuyper, A. & Krings, G. A survey of results on mobile phone datasets analysis EPJ Data Science, Springer, 2015, 4, 10

[10] Bui, N.; Cesana, M.; Hosseini, S. A.; Liao, Q.; Malanchini, I. & Widmer, J. A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques IEEE Communications Surveys & Tutorials, IEEE, 2017

[11] Zhang, Z.; Liu, F.; Zeng, Z. & Beijing, W. Z. A traffic prediction algorithm based on Bayesian spatio-temporal model in cellular network Proc. Int. Symp. Wireless Communication Systems (ISWCS), 2017, 43-48

[12] Yu, P.; Jian-min, W. & Xi-yuan, P. Traffic prediction with reservoir computing for mobile networks Natural Computation, 2009. ICNC'09. Fifth International Conference on, 2009, 2, 464-468

[13] Wang, S.; Guo, J.; Liu, Q. & Peng, X. On-line traffic forecasting of mobile communication system Pervasive Computing Signal Processing and Applications (PCSPA), 2010 First International Conference on, 2010, 97-100

[14] Nikravesh, A. Y.; Ajila, S. A.; Lung, C.-H. & Ding, W. Mobile Network Traffic Prediction Using MLP, MLPWD, and SVM Big Data (BigData Congress), 2016 IEEE International Congress on, 2016, 402-409

[15] Huang, C.-W.; Chiang, C.-T. & Li, Q. A Study of Deep Learning Networks on Mobile Traffic Forecasting

[16] Subrahmanyam, Gowrishankar & Satyanarayana, P.S.. (2012). Neural Network Based Traffic Prediction for Wireless Data Networks. International Journal of Computational Intelligence Systems. December 2008. 379-389

[17] telecomitalia.com/tit/it/bigdatachallenge.html

[18] Zhang, Y., Wang, W., Zhang, S., Fan, D., Kou, B. and Zhou, W., 2016, November. SVR based voice traffic prediction incorporating impact from neighboring cells. In Wireless Personal Multimedia Communications (WPMC), 2016 19th International Symposium on (pp. 142-146). IEEE.