# Nodes Number Estimation based on ML for Multi-operator Unlicensed Band Sharing to Extend Indoor Connectivity

Oluwatobi Baiyekusi*, Haeyoung Lee†, Klaus Moessner‡§
*Department of Computer & Information Sciences, University of Strathclyde, Glasgow, UK
†School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, UK
‡5GIC & 6GIC, Institute for Communication Systems (ICS), University of Surrey, Guildford, UK
§Faculty of Electronics and information Technology, University of Technology Chemnitz, Germany
Email: oluwatobi.baiyekusi@strath.ac.uk, H.Lee@herts.ac.uk, klaus.moessner@etit.tu-chemnitz.de

*Abstract*—Due to ever-increasing data and resource-hungry applications, the needs of new spectrum by mobile networks keep increasing. Unlicensed spectrum is still expected to play a crucial part in meeting the capacity demand for future mobile networks. But if this will be a reality, fair coexistence attained via practical and efficient channel access procedures would be necessary. In designing such channel access schemes, awareness of the number of nodes contending for the channel resource can be strategic. This paper investigates a node number estimation approach using machine learning (ML) techniques. When multiple nodes access the same unlicensed channel, varying idle-time can be associated to a statistical distribution. In this paper, a statistical distribution of the Idle-time slots over the channel are used to characterise and analyse the channel contention based on the number of nodes. Three ML model based approaches are evaluated and the results confirm that the proposed solution's viability but also reveal the best performing ML technique for the task of node number estimations.

*Index Terms*—Coexistence, Unlicensed band, Machine learning, Node Number Estimation.

## I. Introduction

With booming demand for indoor mobile broadband and novel applications such as robots, and Augmented Reality/Virtual Reality (AR/VR) in locations including factories, stadia, airports, and offices, a seamless and high-speed indoor connectivity becomes important [1]. To expand the coverage and extend the connectivity in indoors where the high concentration of mobile traffic is present, mobile operators are interested in innovative solution for faster deployment of indoor networks. While mobile networks are expected to use their allocated spectrum, utilizing sharing schemes (infrastructure/resource sharing) or opportunistic spectrum access schemes, resource expansion would be necessitated due to high demands of indoor wireless services [2]. Hence, due to the benefit of license exemptions, unlicensed spectrum use by mobile operators have been studied widely and the relevant standards was ratified by the 3rd Generation Partnership Project (3GPP) for LTE-LAA [3] and similarly for 5G NR-U [4]. However, fair coexistence remains the area of contention between mobile networks and incumbent technologies. Feasibility studies were conducted at the outset of LTE-LAA's entry into unlicensed bands to evaluate the coexistence impacts on multiple LTE-LAA networks and existing networks operating on same band [5], [6]. One important finding is that the number of nodes operating over the same unlicensed channel can be utilized in designing fair and efficient channel access protocol [6], [7].

In literature, the estimation of node number operating over the unlicensed channel has been investigated [8]–[10]. By using a listen-before-talk (LBT) mechanism, the relationship of the collision probability and the node number was studied in [8] and this study has been utilized as the baseline for related researches. With increase in the node number, estimation of the node number becomes more difficult [9]. The authors in [9] proposed the filter based detection mechanism to increase the accuracy. However, if node number is small, high accuracy is not achieved. In [10], the authors considered the average idle slot interval to propose the estimation technique. The node numbers' variation in the channel was tracked and and a formula of the average idle slots is provided. While the estimation is performed based on the measured average idle slot and the threshold, a large variance was shown in the number of nodes and remains insensitive to smaller increase in the number of nodes on the unlicensed channel.

While above works are non-ML based problems, the algorithms have the limitations to estimate the node numbers which can change either with large or small increase. As a result, to make accurate node number estimations sensitive to small or large increase, in this paper, we study *the node number estimation* mechanisms using *machine learning* [11]. The approach exploits the capability of nodes operating over the unlicensed bands to sense the channel before transmission (i.e. LBT). The periodic but varying idle-time over the channel can be associated to a statistical distribution. The mean and standard deviation of this idle-time distribution can be characterised to the number of active nodes contending over the channel. The dataset acquired from observing the idle-time can be used to train ML models to perform number of nodes estimations operating over unlicensed bands. By using three well-known ML methods, *Multilinear Regression, k-*

*Nearest Neighbour, and Random Forest*, we prove the proposed node number estimation in unlicensed band for multi-operator cellular networks. Using data obtained from ns-3 system simulation, it is shown that not only can node number estimation be performed but with high accuracy regardless of the node number. Awareness of the number of nodes would enable to design the efficient channel access protocols for fair coexistence and extend the performance of indoor connectivity in unlicensed bands.

We adopt the 3GPP indoor scenario for LTE-LAA consisting of multiple mobile networks operating on the same channel (LTE-LAA+LTE-LAA). This is because it has been used widely in literature in evaluating solutions for multiple operators' network coexistence studies in unlicensed bands. It is therefore reasonable to adopt a similar scenario, to evaluate our proposed approach in estimating the number of nodes for existing networks. The remainder of the paper is organized as follows. Section II describes the considered system model and scenario. In Section III, the proposed ML based prediction approach is described. The performance validation is explained in Section IV to show the effectiveness of our proposed approach. Finally, we draw conclusion in Section V.

## II. SYSTEM MODEL

A description of the system model is given in this section including the multi-operator network scenario setup, MAC protocol and data acquisition process.

### A. Network Model

Fig. 1 illustrates our scenario setting, the layout of the network and user devices connected to the network. The network consists of small cells positioned side by side for each network operator. Four cells are operated on each network, aligned and centred along the longer dimension of the building. The separation of the Base Stations (BSs) for each network are uniform across nodes from the same operator. The user devices connected to both networks are randomly positioned within the coverage area inside the indoor environment. The user devices are attached to small cells based on the proximity to the base stations. We consider the downlink transmission but this work could be extended to the uplink.

Similar network parameters are used for both networks such as transmission power for the BS. The indoor scenario
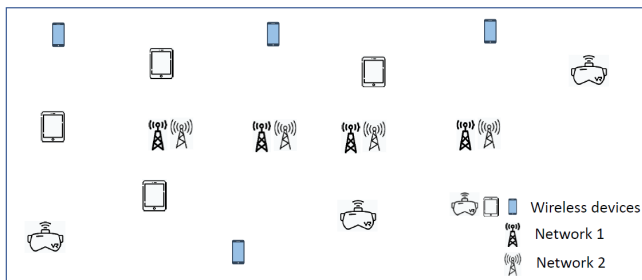


Fig. 1. Network model for mobile network operating in unlicensed bands

considered for the work, adopts the ITU Indoor hotspot (InH) model [12]. For line of sight propagation, which is the case in our network model, (1) is applicable for $10$ m $< d < d_{BP}$ [3] where $d$ denotes the distance between the BS and the user device, while $d_{BP}$ is the breaking point distance.

$$PL(d) = 28.0 + 22\log_{10}(d) + 20\log_{10}(f_c). \quad (1)$$

$PL$ denotes the path loss encountered during transmission and $f_c$ represents the centre frequency of the channel. $PL$ is an important factor which determines the energy detected across all nodes and influences the accurate sensing of transmission and idle-times over the unlicensed channel.

### B. LTE-LAA MAC Model

The LTE-LAA channel access procedure operates the LBT scheme which requires a node to sense and detect any ongoing activity over the channel before transmission is initiated. In the channel access procedures documented, only the downlink transmission is specified [13]. A carrier selection is performed through which the physical downlink shared channel (PDSCH) can transmit data to user devices. It then follows an arbitration process whereby firstly, a set time called the defer duration $T_d$, when the channel must be sensed to be idle must be observed. Progressing, it selects a random number $N$ from a range of contention window (CW) using a uniform distribution. These CWs are in different sizes depending on the priority class decided by the traffic type, which are given in Table I. Specifically, the value of CW begins with the smallest CW size and increments to the next higher value when 80% HARQ-ACK is detected as NACKs in the reference subframe. Each priority class has a min & max CW size ($CW_{\min}$ & $CW_{\max}$) and their maximum respective channel occupancy time ($T_{cot}$) for every transmission opportunity to any user device. Once the random number $N$ is chosen for CW and the channel is still idle, is decremented by $N-1$ after each slot period $T_{sl}$. On reaching a value of $0$, transmission begins with the specified $T_{cot}$. $T_d$ and $T_{sl}$ are set times at 16 $\mu s$ and 9 $\mu s$, respectively.

TABLE I
CHANNEL ACCESS PRIORITY CLASS FOR LTE-LAA [13]

| Priority Class | $CW_{\min}$ | $CW_{\max}$ | $T_{cot}$ | Allowed $CW$ sizes |
|---|---|---|---|---|
| 1 | 3 | 7 | 2 ms | 3, 7 |
| 2 | 7 | 15 | 3 ms | 7, 15 |
| 3 | 15 | 63 | 8 or 10 ms | 15, 31, 63 |
| 4 | 15 | 1023 | 8 or 10 ms | 15, 31, 63, 127, 255, 511, 1023 |

### C. Data Obtainment

In order to estimate the node number, obtaining data on the unlicensed channel usage is crucial. As all contending nodes and devices sense the activity of the channel, idle-time intervals over the channel constitutes the relevant data to addressing the estimation task. This means the data is obtained as the nodes perform their LBT operations, without any signalling
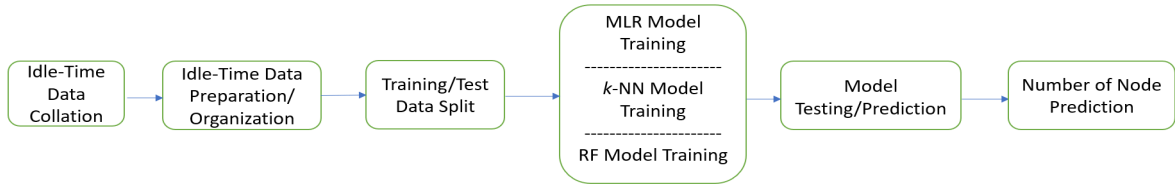
Fig. 2. The operational procedures of proposed ML-based number of node estimation

between nodes. This data is usable for the purpose described in this paper because the difference between the randomly selected backoff number $N_i$ for each contending node or device represents a new distribution known as the uniform difference distribution. This new distribution represents the relationship between all nodes actively contending over the channel. Features of this distribution can then be analysed and classified to aid prediction and estimation of the number of nodes. The time units of the idle-time intervals are slots, as the backoff counter is decremented after a slot time $T_{sl}$ (9 $\mu s$). A mathematical representation of the uniform difference distribution can be represented by the equation below:

$$D = N_1 - N_2 - \cdots - N_k, \tag{2}$$

where $N_i$ are independent variables as the node $i$'s CW value, operating within a set CW range and selected via a uniform distribution. $D$ stands for the new uniform difference distribution which constitutes the idle-time intervals used.

Extensive simulation campaigns are carried out for different number of nodes and different randomly located nodes. A collation of the idle-time intervals associated with their respective nodes numbers are saved on network log files in the simulator. The number of nodes chosen as part of the evaluation study are $4, 8, 10, 16, 20, 24, 30, 36, 40$.

Idle-time information in ns-3 is represented in microseconds. Then, these are converted to the number of slots by using the following relation.

$$N_{sl} = \frac{T_{idle}}{T_{sl}}, \tag{3}$$

where $T_{idle}$ denotes idle-time within each transmission interval. $N_{sl}$ represents the number of slots within the transmission interval. $N_{sl}$ for every idle-time within the simulation time provides the dataset used to train the ML models for node number estimations. For each node number, an average of over $18,000$ idle-time measurements are carried out where collated for each simulation performs. The process of our node number estimation mechanism is depicted in Fig. 2.

## III. PROPOSED MACHINE LEARNING ESTIMATION DESIGN

In this section, we present three machine learning models developed to perform the node number estimation. This task essentially involves data categorization. In our scenario, multiple categories are formed from the data obtained from channel sensing which indicates the number of nodes. Hence *Multiclass Classification* ML techniques are used to train the estimation/prediction models. Features are selected and

categorized which form the labels to be used by the Multiclass Classification models. The idle-time intervals represent a stochastic distribution, and the mean and standard deviation of features from the dataset constitute the data points used in training the ML models. Combining the mean and standard deviation of each feature sets, allows grouping and association with specific number of nodes (labels).

### A. Data Preprocessing and Organization

The dataset obtained during simulation, was cleaned by removing outliers and errors in the data. A quick histogram plot revealed a convoluted distribution of the cleaned data, which is consistent with the shape of a data plot from a uniform difference distribution. Thereafter, the dataset were subdivided into smaller portions, which constitute the features. These subsets (features) of the full dataset all maintained the convoluted shape of the uniform difference distribution. The mean and standard deviation of these subsets became the data points for training the models. The same features were used to train all three ML techniques considered for evaluation. For organizing the data, a matrix structure was adopted such that $\mathbf{X} = [X_1 \cdots X_N]_{M \times N}$ and $X_n = [x_{1n} \cdots x_{Mn}]^T$; $M$ is the number of datapoints and $N$ is the number of the features.

### B. Multilinear Regression

Multilinear regression is used for classification problems involving multiple independent variables and a dependent variable. The multiple input variables influence the output variable, which in our case are the mean and standard deviation (input variables) and the number of nodes (output variables) respectively. Achieving a best fit for the lines of regression is important, as they influence the accuracy of the prediction. The Multilinear model is trained using (4) where $Q$ is the dependent variable and predicted output. $c$ is the intercept of the line of regression, $\alpha_1$ and $\alpha_2$ are the regression coefficients, with $X_1$ and $X_2$ being the input matrices, where $X_1$ and $X_2$ indicate mean and standard deviation features of datapoints, respectively.

$$Q = c + \alpha_1 X_1 + \alpha_2 X_2 \tag{4}$$

### C. k-Nearest Neighbour (k-NN)

The $k$-NN algorithm utilizes a different approach to ML. It is a non-parametric supervised learning algorithm, working on the assumption that data points in close proximity to the input data fall under the same group. Our model is trained on labelled data consisting of the number of nodes. In making

| Parameter | LTE-LAA |
|---|---|
| Slot time | 9 $\mu$s |
| Defer Time | 43 $\mu$s |
| Tx Power BSs | 18 dBm |
| Carrier Frequency | 5 GHz |
| Bandwidth | 20 MHz |
| Total Subcarriers | 1200 |
| CW$_{min}$ & CW$_{max}$ | 15 & 1023 |

a class label prediction to input data, $k$ (the number of neighbours) is crucial to the accuracy of the model and needs to be provided to the classifier, through which the classification is made. The Euclidean distance method in (5), most common for real-valued points, is used to calculate the distance metric between data points.

$$d(\mathbf{x}_i, \mathbf{Y}) = \sqrt{\sum_{j=1}^{2}(x_{ij} - y_j)^2} \qquad (5)$$

$\mathbf{x}_i$ are the training datapoints and Y is the input data from each test data through which the euclidean distance is calculated. This distance determines the nearest neighbour. The $k$ nearest neighbours are chosen, and the number of data points in each categories is counted. The input data will be assigned to the category for which the number of the neighbour is maximum.

### D. Random Forest (RF)

RF can be also used for classification tasks. As a hierarchical decision making approach, RF possesses a structure which constructs an ensemble of decision tress. The decision trees contain branches which aid the splitting of data into subsets. For multiple decision trees forming an ensemble, they can predict more accurate results, particularly when the individual trees are uncorrelated with each other. The output from individual decision trees are aggregated to make a final classification prediction.

## IV. RESULT DISCUSSION

In this section, results are presented showing the performance of the three ML models trained on the channel idle-time dataset acquired for LTE-LAA coexistence from the simulation campaign on ns-3. Simulation parameters are given in Table II.

For training, testing and evaluating our ML based node number estimation models, we use python libraries such as Pandas, Numpy, Sklearn and Yellowbrick. The collected dataset is split to 70% and 30% for model training and testing, respectively. We evaluate the performance of the ML models for two level of granularity of node numbers: Coarse and Fine estimation. The first case is trained for Coarse Estimation (CE) with a group of dataset of 10, 20, 30, 40 nodes (i.e. idle-interval distribution data when the gap of node number is 10). The second case is for Fine Estimation (FE) with the data of 4, 8, 10, 16, 20, 24, 30, 36, and 40 nodes (when the gap of

node number is 4). We compare the performance of both cases having different granularity.

The result in Table III reveals a significantly high negative correlation for CE. This is expected as the level of granularity is lower i.e. CE, and the distinctions between the features are more detectable. However, a lower but nonetheless sufficient correlation coefficient is observed for FE. This shows a relationship exists between the mean and standard deviation which are the chosen features of the dataset.

Beyond the correlation coefficient, as the performance indicators of three ML based estimation methods, we calculate the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the R-squared. These three metrics are widely used for evaluating Multilinear Regression models. For the $k$-NN and RF estimation models, the F1 score and the classification prediction error metric are used. The F1 score is a harmonic mean of the precision and the recall metrics while the classification prediction error metric shows the quantity of errors in classification.

### A. Multilinear Regression Results

In Table IV, the Multilinear regression results i.e. MAE, RMSE and R-squared value are presented. For CE, it is clearly shown that fairly accurate predictions can be achieved via Multilinear regression. The R-squared value indicates the degree of variance in the data (goodness of fit) and scores high at 0.9057 out of 1.0 for CE. This reflects the correlation earlier measured that confirms the relationship between the features and the labels for classification. On the other hand, a R-squared of 0.5013 is measured in the prediction for FE. A relatively higher MAE, RMSE and consequently lower R-squared are recorded for FE. This is expected as the number of nodes increases due to the less linear relationship between the features. The nominal results for FE for Multilinear regression and high score for CE, shows Multilinear regression can be considered for node number estimation depending on the level of accuracy desired.

### B. k-Nearest Neighbour Results

First, the F1 score is illustrated in Fig. 3. For the case of coarse estimation, a F1 score is measured as 1.0 regardless of the value of $k$ (for $k = 1, 5, 10, 20$). This means no estimation error is made for this category estimation. For fine estimation, it is shown the measured F1 scores are fluctuated for different $k$ and the number of nodes. It is interpreted that the smaller gap of features (the gap of node number is 4) leads to error in estimation across different $k$ and different node numbers. The similar results are observed in Fig. 4 and 5 depicting the class

TABLE III
CORRELATION OUTPUT FOR DATASET FEATURES

| Category | No. of Nodes | Mean | Standard Deviation |
|---|---|---|---|
| CE | 10-20-30-40 | -0.92345 | -0.93856 |
| FE | 4-8-10-16-20-24-30-36-40 | -0.65204 | -0.67836 |

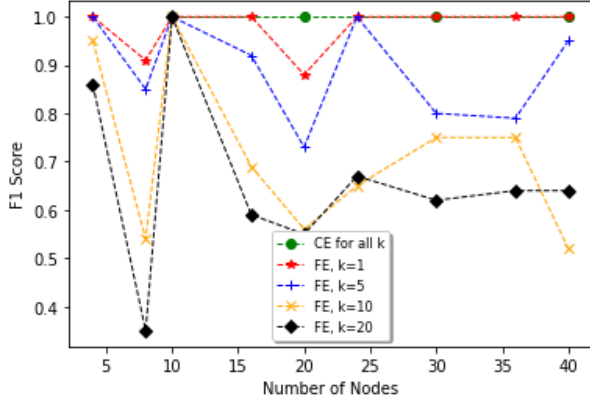| Node Number Category | MAE | RMSE | R-Squared |
|---|---|---|---|
| 10-20-30-40 (CE) | 3.1539 | 3.6961 | 0.9057 |
| 4-8-10-16-20-24-30-36-40 (FE) | 7.4723 | 8.7169 | 0.5013 |



Fig. 3. $k$-NN F1 performance based on different $k$ values for Coarse and Fine Estimation

prediction error. In Fig 4 for $k = 1$, no wrong classification is observed. For all class of node number, the right class could be mapped as a result of estimation. However, for other $k$ ($k = 5, 10, 20$), classification results including estimation error are observed. Fig. 5 shows the result of $k = 20$ (due to space limitation, only the result of $k = 20$ is included). For instance, no estimation error is made for the case of 10 node, but for the case of node number $40$, wrong number is sometimes estimated as $16, 24, 30$. The plot in Fig. 5 are particularly helpful in identifying which labels are problematic for the model when particular node numbers are being estimated.

## C. Random Forest Results

Fig. 6 presents the F1 score obtained by RF based estimation. Similar to the $k$-NN based estimation, the high F1 score
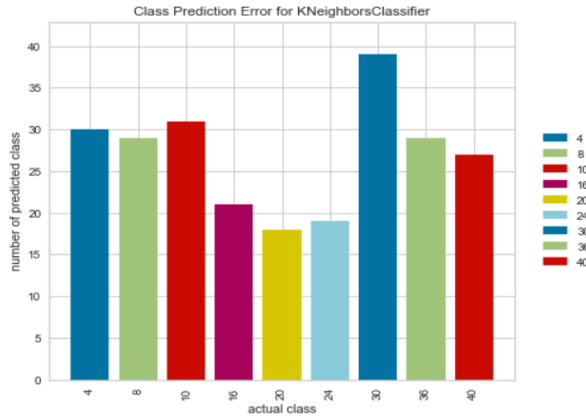


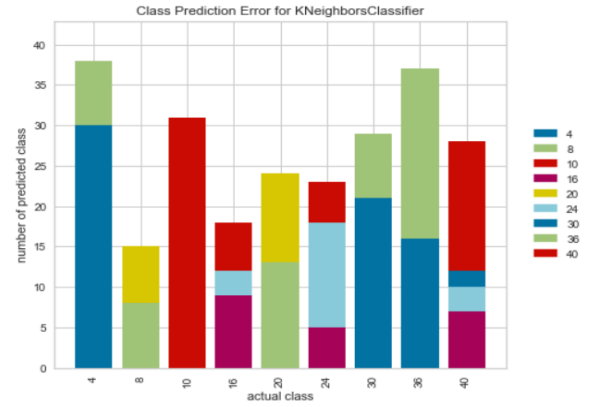Fig. 4. $k$-NN class prediction error for $k = 1$ for Fine Estimation



Fig. 5. $k$-NN class prediction error for $k = 20$ for Fine Estimation
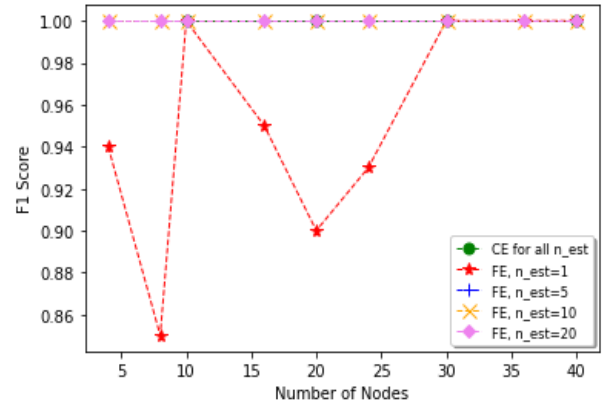


Fig. 6. Random Forest performance for different $n$ estimators for Fine Estimation

1 is achieved for coarse estimation over all node numbers. However, a distinct performance improvement is observed for fine estimation when compared with the $k$-NN estimation. We would expect with one estimator, that not high performance can be achieved bearing in mind the nature of RF models. However, for 5, 10, 20 estimators, a consistent F1 score of 1 is measured. We see a similar trend with the class prediction error plot in Fig 7, that shows some wrong classification when number of estimators = 1. However, accurate estimations are observed when higher estimators are used in the model as shown in Fig 8.

## D. Further Analysis

We compare the performance of all three ML methods evaluated in this paper. From the results presented, we observe the $k$-NN and RF ML techniques offer better performance especially for fine estimation when compared to Multilinear regression. This can be explained due to the Multilinear regression approach which requires a good fit for the regression line which can sometimes be challenging depending on the data points. However, when $k$-NN is compared with RF, we find RF performs better. This is because the RF approach seeks to make clear distinctions between uncorrelation within
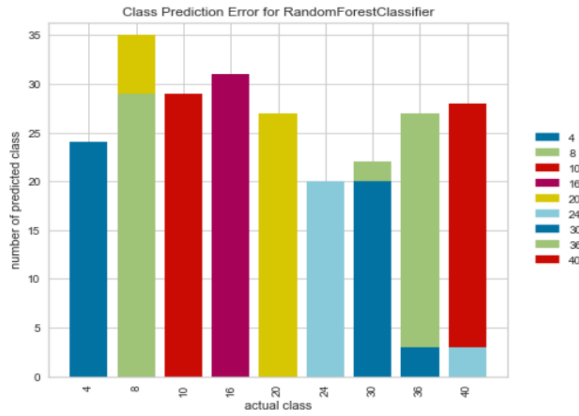
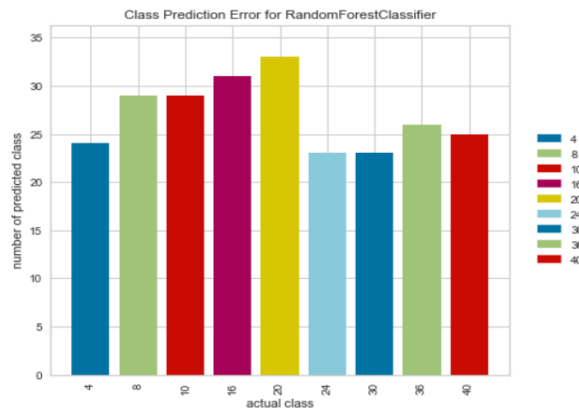Fig. 7. Random Forest class prediction error for $n = 1$ estimators for Fine Estimation



Fig. 8. Random Forest class prediction error for $n = 5$, 10 & 20 estimators for Fine Estimation

the dataset. This technique is more efficient for classification compared to $k$-NN which classifies based on proximity. Also very low $k$ value could cause overfitting to the model which negatively influences the prediction, while too high $k$ value can lead to underfitting and higher computational cost in calculating the distance for all the points. RF on the otherhand has a lesser training time when compared to the other ML models and the risk of overfitting is significantly reduced due to the use of multiple trees. The low correlation between each decision trees actually produces more accurate predictions. For real network implementations, it will be best to use the lowest number of RF estimators that achieves the high accuracy needed to train the model. This will reduce both training and classification time.

## V. Conclusions

In this paper, our effort concentrated on the evaluation of three well used ML techniques for multiclass classification. These were used to predict number of nodes based on pre-processed idle-time input data supplied to the trained ML models under a multi-operator mobile network scenario in unlicensed bands. The $k$-NN and RF models outperform the

Multilinear regression model. Furthermore, the RF models reveals a distinct performance above them. The work presented in this paper, clearly shows the validity of the method adopted and accuracy of estimation based on the data supplied to train the ML models. Our approach will be extended to design channel access procedures which can contribute to the implementation of fair spectrum use in unlicensed bands, which will form part of future work.

## Acknowledgment

## References

[1] Dedicat 6G, "Deliverable 2.3: Dynamic coverage Extension and Distributed Intelligence for human Centric Applications with assured security, privacy and Trust: from 5G to 6G," Tech. Rep., Feb. 2022.

[2] G. Gür, "Expansive networks: Exploiting spectrum sharing for capacity boost and 6G vision," *J. Commun. Netw.*, vol. 22, no. 6, pp. 444–454, 2020.

[3] 3GPP TR 36.889, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Licensed-Assisted Access to Unlicensed Spectrum (Rel. 13)," Tech. Rep., v.13.0.0, 2015.

[4] ——, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on NR-based access to unlicensed spectrum (Rel. 16)," Tech. Rep., v.16.0.0, 2018.

[5] A. Mukherjee, J.-F. Cheng, S. Falahati, H. Koorapaty, D. H. Kang, R. Karaki, L. Falconetti, and D. Larsson, "Licensed-Assisted Access LTE: coexistence with IEEE 802.11 and the evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 50–57, June 2016.

[6] M. Mehrnoush, S. Roy, V. Sathya, and M. Ghosh, "On the Fairness of Wi-Fi and LTE-LAA Coexistence," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 735–748, 2018.

[7] C. Cano and D. J. Leith, "Coexistence of WiFi and LTE in unlicensed bands: A proportional fair allocation scheme," in *Proc. IEEE ICC workshop*, 2015, pp. 2288–2293.

[8] H. Ko, J. Lee, and S. Pack, "A Fair Listen-Before-Talk Algorithm for Coexistence of LTE-U and WLAN," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10 116–10 120, Feb. 2016.

[9] G. Bianchi and I. Tinnirello, "Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network," in *Proc. IEEE INFOCOM*, vol. 2, Apr. 2003.

[10] S. Chun, D. Xianhua, L. Pingyuan, and Z. Han, "Adaptive Access Mechanism with Optimal Contention Window Based on Node Number Estimation Using Multiple Thresholds," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2046–2055, Apr. 2012.

[11] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions," *IEEE Access*, vol. 7, pp. 137 184–137 206, 2019.

[12] ITU-R M.2135, "Guidelines for evaluation of radio interface technologies for IMT-Advanced," Tech. Rep., June v.1, 2009.

[13] 3GPP TR 36.213, "3rd Generation Partnership Project (3GPP), "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Rel. 14)," Tech. Rep., v.14.2.0, 2019.