ML-Aided Dynamic Clustering and Classification of UEs as VBs in D2D Communication Networks

Conference Paper · April 2023

CITATIONS
0

READS
22

8 authors, including:

Iacovos Ioannou
University of Cyprus
21 PUBLICATIONS **81** CITATIONS
SEE PROFILE

Ala' Khalifeh
German Jordanian University
140 PUBLICATIONS **1,217** CITATIONS
SEE PROFILE

Christophoros Christophorou
University of Cyprus
55 PUBLICATIONS **333** CITATIONS
SEE PROFILE

Vasos Vassiliou
University of Cyprus
141 PUBLICATIONS **1,668** CITATIONS
SEE PROFILE

ML-Aided Dynamic Clustering and Classification of UEs as VBs in D2D Communication Networks

Iacovos Ioannou*†, Prabagarane Nagaradjane‡, Ala' Khalifeh¶, Christophoros Christophorou*†, Vasos Vassiliou*†, Gundepudi V. Surya Sashank‡, Charu Jain‡, Andreas Pitsillides*§
* Department of Computer Science, University of Cyprus
† CYENS - Centre of Excellence, Cyprus
‡ Department of ECE, Sri Sivasubramaniya Nadar College of Engineering Chennai, India
§ Department of Electrical & Electronic Engineering Science University of Johannesburg (Visiting Professor)
¶ Department of Electrical Engineering and Information Technology, German Jordanian University

***Abstract*—With the next generation of mobile devices and streaming services such as Virtual Reality, Augmented Reality, and Meta being available worldwide, the network's data rate, latency, and connectivity must be improved. Even though 5G provides the user with the required Quality of Service (QoS) in terms of data rate and reduced delays by transmitting signals in the higher frequencies (called New Radio (NR)), it faces a lot of attenuation, leading to a short communication range. However, to meet goals set by utilising 6G requirements, many technological advancements and components must be incorporated into the network. The dense disposition of small cells will help reduce the network traffic in hotspot areas and increase the coverage and spectral efficiency. Nonetheless, the current deployment of 5G Base Stations (BSs) and small cells is static and cannot move around even though they are deployed in hot spot areas, leading to high operational costs. Furthermore, more than these static deployments of base stations can be required in an unpredictable scenario of extreme crowd movement. To overcome these issues, Device to Device (D2D) Communication with the dynamic deployment of Virtual Base stations (VBSs) can be called upon, which can be achieved by using User Equipment (UE) such as phones or laptops to mimic the functions of a Base Station (BS). Therefore, in this paper, a User Equipment based Virtual Base Station (UE-VBS) is studied, which will act as a secondary base station and, in turn, help alleviate the traffic load in the network. Specifically, as one UE cannot relieve the entire network traffic load, the network area is split into different clusters by using an unsupervised Machine Learning (ML) clustering technique(i.e., K-Means with Mean Shift Clustering), and a single UE is selected to act as a VBS for that cluster with the utilisation of supervised ML classification techniques (i.e., Decision Trees, Logistic Regression, Linear Discriminant Analysis And Quadratic Discriminant Analysis, Linear Support Vector). In our work, we utilise the K-means along with mean shift clustering techniques to cluster simulated network areas accurately. Also, we use and compare different classification machine learning techniques to predict/classify whether user equipment can be employed as a VBS and become UE-VBs. Our simulation study reveals that the Decision Tree algorithm achieves the highest accuracy in categorising the eligible UEs as UE-VBs.**

***Index Terms*—D2D Communication, VBs, UE-VBs, VBs Classification to UE-VBs, K-Means Clustering, Mean Shift Clustering, Machine Learning, 5G, 6G, D2D, Decision Trees, Logistic Regression, Linear Discriminant Analysis And Quadratic Discriminant Analysis, Linear Support Vector**

I. INTRODUCTION

To satisfy the highly demanding network application requirements of a UE device that is accessing the internet, such as high-quality video streaming, fast gaming, augmented reality, etc., the data rate, latency, and connectivity must be improved. To facilitate the needs mentioned above, the existing 5G mobile network has to be upgraded to the next generation, i.e., 6G, since it uses signals with higher frequencies. These signals will face a lot of attenuation, which will finally lead to a short communication range. Thus, the current 5G mobile generation will require several critical key technological components to meet these ambitious goals of 6G, including even more densification with heterogeneous networks and small cells. The dense deployment of small cells will play a significant role as they can be set up more specifically to relieve traffic in crowded areas and increase coverage and spectral efficiency.

However, the existing deployment of base stations and small cells is static and generally/regularly deployed on demand in hotspot areas. In the case of unpredictable population mobility, these static deployments of BS can be ineffective and costly in terms of both capital and operational expenses. Device-to-device (D2D) communication can contribute to their reduction with the utilisation of Users' Equipment. By providing special packets for the participating users, telecoms can spare the usage of small cells and the hardware needed to establish the small cells. D2D communication enables adjacent user equipment (UE) to bypass the base station and establish direct links to either share their connection and act as relay stations or directly communicate and exchange data. D2D can operate in both licensed and unlicensed spectrum and is generally transparent to the cellular network. [1], [2].

This research is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°739578 and the government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination, and Development.

979-8-3503-0045-1/23/$31.00 ©2023 IEEE 2

Additionally, along with the D2D communication scheme, two things are to be achieved: Dynamic deployment of BS and creation of a cost-efficient BS. This can be realised by making a set of UE, such as phones, laptops, etc., emulate the actions of a BS and act as a D2D-relay device that shares its bandwidth and all the services of a BS [2], [3]. Hence, the UE can be termed a VBS. Our research focuses on the scenario where UE-based VBs can be dynamically selected among the UEs available in the network on an at-will basis to deal effectively with the nonstationary and non-uniform mobile traffic distribution concerning space and time domain. Thus, D2D communication is enabled. Since one UE cannot provide enough micro-segmentation for a macro cell, the cell has to be split into several small cells, with one UE acting as the UE-VBS in each cell (as shown in the Fig. 1). We propose an efficient clustering technique to implement this concept: Mean shift clustering to find the number of clusters and K-Means Clustering to segregate the UEs into small cells. Communication parameters such as received power, battery discharge rate, and the Signal-to-Interference Noise Ratio (SINR) are used to select the best set of UE-VBSs to supplement an overlay-loaded network.

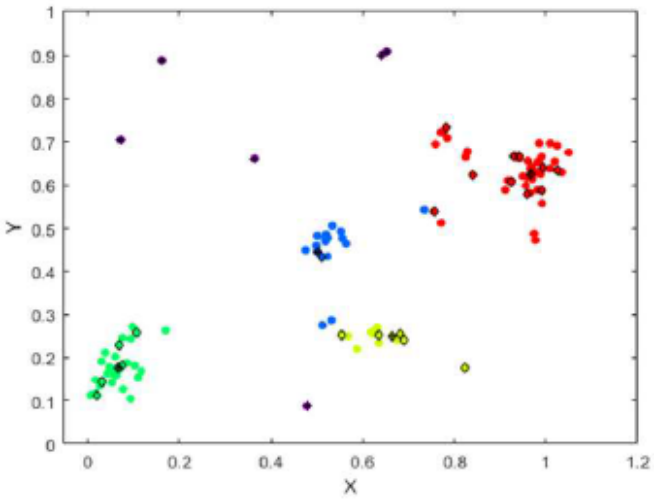

Fig. 1. Selection of active UE-VBSs using affinity propagation clustering [1]

The novelty of this paper is that it executes a two-step Machine Learning (ML) approach to break the cell under a cell into multiple clusters (micro-segmentation over a cell) and to select the most appropriate devices to be the UE-VBs under each cluster. Firstly, it employs the results of UEs' separation into clusters using mean shift clustering to determine the number of clusters and the K-Means clustering algorithm as a clustering technique for selecting; secondly, the most suitable UE device to be UE-VBs using ML classification techniques. More specifically, it is the only approach that evaluates the devices to non-UE-VBs and UE-VBs devices according to specific features and uses ML approaches in order to select the most appropriate device to be UE-VBs after clustering results. Thus, the key contributions of this paper are summarised below:

- The approach uses two steps, clustering and classification, to determine the most suitable UE device to become UE-VBS.
- We use mean shift clustering to determine the number of clusters and the K-Means clustering algorithm for separating the UEs into clusters in a footprint.
- This work proposes an ML model to classify the potential UE-VBs devices to VBs and non-VBs with the utilisation of specific metrics (i.e., data rate, Signal-to-Noise Ratio (SNR), Quality of Service or QoS, battery capacity, storage, battery level, CPU name/model, CPU speed and RAM) with the use of Decision Trees, Logistic Regression, Linear Discriminant Analysis And Quadratic Discriminant Analysis and Linear Support Vector algorithms.
- It compares the classification approaches used (Decision Trees, Logistic Regression, Linear Discriminant Analysis And Quadratic Discriminant Analysis, and Linear Support Vector) and concludes that Decision Trees achieve better accuracy than the others.
- The proposed model selects the most appropriate UEs as UE-VBs using the score of the metrics after label encoding. The UE accreditable result (i.e., YES, NO) with the highest score will be deemed as the UE-VB of a specific cluster.

Following this introductory section, we structure the rest of the paper as follows. Section II provides a review of related research works using ML for clustering UE devices under a cell at a D2D communication network and the background information needed for each approach. Thus, it provides a brief description of the investigated approaches. Section III elaborates on the problem that this paper focuses on tackling and highlights the importance of clustering and classifying in a D2D communication environment. In Section IV, we show the methodology used in our approach. More specifically, we describe the datasets used, the hyper-parameters used, how the validation of our ML models is executed related to the units used, and we show how the classification parameter is added to the dataset. Section V provides some brief information about the clustering execution approach used by the papers classification approach. Next, Section VI shows the classification of ML approaches considered for investigation and presents the performance evaluation and discussions on the results achieved from the experiments conducted. Finally, Section VII draws our conclusions and future work.

II. LITERATURE REVIEW AND BACKGROUND INFORMATION

This section provides the literature review on clustering devices under a D2D communication network along with the required background information related to our investigation.

A. LITERATURE REVIEW

This section provides a literature review on clustering devices under a D2D communication network. According to our

research, classification approaches to classify UE-VBs do not exists in the open literature.

P. Swain et al. [3] focused on how the UE-VBSs can be selected dynamically among the available UEs to deal with the non-uniform distribution of mobile traffic. Specifically, the authors proposed an efficient clustering technique (Affinity Propagation Clustering) to select the best UE-VBSs to supplement highly loaded Small Base Stations (SBSs). The algorithm was implemented and evaluated in MATLAB and validated using the NS3 simulator. X. Wang et al. [4] proposed the concept of a VBS, dynamically formed for each cell by assigning virtualised network resources. The authors proposed novel energy-saving schemes exploiting Virtual Function (VF) for the network planning and traffic engineering stages. Z. Zhu et al. [5] introduced the Time Division Duplex (TDD) WiMAX Software Defined Radio (SDR) BS implemented on a commodity server in combination with a Remote Radio Head (RRH) design. He also presented the first working prototype of a VBS pool. The results from this pool prototype for WiMAX verified that the solutions could meet system requirements, including jitter, synchronisation, and latency. I. Hwang et al. [6] presented a holistic view on hyperdense HetSNets, including fundamental preference in future wireless systems, technical challenges, and also the recent technological breakthroughs made in networks such as these. C. Christophorou et al. [7] proposed the Cella Ecosystem (CelEc) concept for the emerging 5G ultra-dense Networks, to supplement and evolve the Mobile Radio Network by using UEs. C. B. Lucasius et al. [8] proposed a novel approach to the problem of k-medoid clustering of large data sets using a genetic algorithm. Genetic algorithms comprise a family of optimisation methods based on principles of natural evolution. L. Gavrilovska et al. [9] elaborated on the 5G-related topics while identifying the key challenges for future research in 5G standardisation activities. T. Pfeiffer et al. [10] proposed a new fronthaul functional division that could improve the bit-rate requirements by transporting baseband signals instead of sampled radio waveforms. Z. Kong et al. [11] designed a distributed resource management middleware to realise a resource management strategy for Baseband Unit (BBU) clusters to improve energy efficiency.

B. BACKGROUND INFORMATION

This subsection provides the background information needed for the clustering and classification approaches.

1) Clustering: This section provides a brief description related to the K-Means and means shift clustering approaches.

a) K-Means Clustering Algorithm: The K-Means clustering algorithm is a technique for unsupervised machine learning to find data object groupings inside a dataset. In this case, the data objects are UEs, and the data set is location data. K-Means is among the first and most accessible clustering methods. UEs must also be clustered into non-overlapping groups, which K-Means can accomplish. K-Means clustering algorithm uses Euclidean distance as the metric and cluster count as the parameter [12]. The optimal number of clusters (K), which is the only parameter examined by the K-Means clustering algorithm, is of the utmost importance. This is determinable using the mean shift clustering approach.

b) Mean Shift Clustering: Mean shift is a clustering algorithm that assigns data points to clusters iteratively by shifting points toward the mean (highest density of data points in the region), also referred to as the mode-seeking algorithm. In mean shift, the algorithm determines the number of clusters based on the data. Mean-shift is founded on Gaussian Kernel Density Estimation (KDE). It is a technique for estimating a given data set's underlying distribution, commonly known as the probability density function. It operates by applying a kernel to each data point. The kernel is a common weighting function used in the convolution process [13].

2) Classification: This section will briefly introduce our classification approaches (i.e., Decision Tree, Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Linear Support Vector Classification (SVC)).

a) Decision Tree: A decision tree uses a root node, which does not contain any incoming branches. The outward branches from the root node feed into the internal nodes, also known as decision nodes. The leaf nodes represent all the potential outcomes of the dataset. Decision tree learning implements a divide-and-conquer strategy by employing a greedy search to identify the optimal split points inside a tree. This process is then repeated in a recursive, top-down manner until all or the great majority of items have been classified under predetermined class labels. Noting that pruning is often used to minimise complexity and prevent overfitting (which makes decision trees complex), pruning is the process of deleting branches that split into low-value attributes. The CART algorithm (which stands for "classification and regression trees") uses Gini impureness to choose the ideal attribute to split on. Gini impureness estimates how frequently a randomly selected attribute is incorrectly classified. A lower value is preferred when utilising Gini impurity for evaluation. Gini impurity is the probability of misclassifying a random data point depending on the dataset's class distribution. [14].

b) Logistic Regression: Logistic regression is a statistical technique used to construct machine learning models using dichotomous, or binary, dependent variables. Logistic regression characterises data and the connection between a dependent variable and one or more independent variables. The nature of independent variables may be nominal, ordinal, or interval. The term "logistic regression" is derived from the concept of the logistic function it employs. Commonly, the logistic function is also known as the sigmoid function. The result returned by the logistic function is between zero and one. Note that the logistic function is at the heart of the statistical technique known as logistic regression. As with linear regression, the representation of logistic regression is an equation. The logistic regression algorithm (in the equation) must be determined from training data. This is done via maximum-likelihood estimation. Maximum-likelihood estimation makes assumptions about the data's distribution [15].

c) Linear Discriminant Analysis And Quadratic Discriminant Analysis: The discriminant analysis methods can be utilised for classification and dimension reduction. LDA is widely utilised since it is both a classifier and a dimensionality reduction technique. Quadratic Discriminant Analysis (QDA) is a variant of LDA that permits nonlinear data separation. LDA/QDA are categorization and dimension reduction approaches that can be regarded from two perspectives. The probabilistic first interpretation facilitates comprehension of the LDA/QDA's assumptions. From a probabilistic standpoint, both utilise Bayes' rule to compute the posterior probability that an observation x belongs to class k; as a result of the normal assumption, the posterior is characterised by a multivariate Gaussian with the same expected covariance matrix for all classes. New points are categorised by computing the discriminant function (the posterior probability enumerator) and returning the biggest class, k. Eigendecomposition of the within-class and between-class variances can be used to identify the discriminant variables. Fisher describes this as an approach for lowering dimensionality in which each successive transformation is orthogonal and maximises the between-class variance compared to the within-class variation. This method converts the feature space to an affine space with K-1 dimensions. After sphering the input data, new points can be categorised by identifying the nearest centroid in affine space while taking class priors into consideration [16].

d) Linear Support Vector Classification: SVM chooses the extreme points and vectors that help to create the hyperplane. These extreme cases are called "support vectors"; hence, the support algorithm is termed a "Support Vector Machine". Using a decision boundary or hyperplane, two different categories are identified. A hyperplane in n-dimensional space can have many lines or decision boundaries to separate classes. However, we must identify the optimal decision boundary for classifying the data points. The optimal boundary is known as the SVM hyperplane. The dimensions of the hyperplane depend on the dataset's features; therefore, if there are two features, the hyperplane will be a straight line. Note that the SVM always generates a hyperplane with a maximum margin or the maximum distance between data points. The support vectors are the data points or vectors nearest to the hyperplane; they influence the hyperplane's position. Because they support the hyperplane, these vectors are known as support vectors. Therefore, the SVM algorithm aids in locating the optimal line or decision boundary. The SVM algorithm identifies the nearest point between two classes' lines. These are known as support vectors. The margin is the distance between the vectors and the hyperplane, and the objective of SVM is to maximise this margin. The hyperplane with the greatest margin is known as the ideal hyperplane. Linear SVM is used for linearly separable data, i.e., data that can be classified into two classes using a single straight line [17].

III. PROBLEM DESCRIPTION

This section provides the problem description that our approach tries to tackle. Mobile phones were mostly the only device expected to be supported in conventional cellular networks (as shown in Fig. 2). With the exhaustive use of the Internet and its multiple applications, the problem of handling various classes of traffic to meet the different Quality of Service (QoS) requirements of diverse applications like video streaming, data, etc., resulted in the traffic to cropped up according to the QoS and Quality of Experience (QoE). A similar situation is arising now with the requirement to support several devices and applications with varying QoS requirements to provide a better experience. Unlike earlier generations of cellular networks, 5G and 6G networks are anticipated to enable gadgets such as smartwatches, driverless vehicles, the Internet of Things (IoT), and tactile Internet. In particular, mobile broadband services, large machine-type communications, and ultra-reliable, low-latency communications would be provided in 5G and 6G, per the International Telecommunication Union (ITU). The many types of devices require more complicated and sophisticated networks that can handle high throughput while providing low latency in data delivery, an efficient energy consumption strategy, high scalability to accommodate a large number of devices, and ubiquitous connectivity for users [2]. These requirements are outlined, followed by an explanation of how potential solutions can meet them are explained below [17].

5G and 6G cellular network is envisioned to support devices like smartwatches, autonomous vehicles, the Internet of Things (IoT), and tactile Internet, unlike previous generations of cellular networks. In particular, according to the International Telecommunication Union (ITU), the three types of service scenarios to be supported in 5G are Enhanced Mobile BroadBand (EMBB) services, Massive Machine-Type Communications (MMTC), and Ultra-Reliable Low-Latency Communications (URLLC). The various types of devices need more complex and sophisticated networks to support high throughput while providing low latency in data delivery, efficient energy consumption scheme, high scalability to accommodate many devices, and ubiquitous connectivity for users [2]. These requirements are described, and how potential solutions can meet them are explained below [17].

IV. METHODOLOGY

Having understood the previous experiments conducted concerning this problem statement, the next step was to deal with the methodology of clustering implementation. We used the findings of mean shift clustering to find the number of clusters and the K-Means clustering algorithm to do the classification of the devices as VBS accreditable or non-VBS accreditable in conjunction with ML approaches (i.e., Decision Tree, Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis and Linear Support Vector Classification). The flowchart of implementation is shown in Figure 3. The machine learning problem faced here was binary classification since the problem was to predict/classify whether a user's equipment could be a virtual base station. Therefore the entire methodology was split into three parts: i) Data selection which includes data engineering; ii) Units used for the training of the

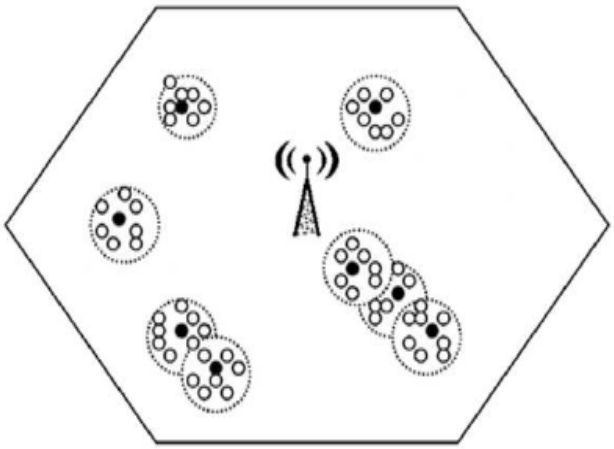

Fig. 2. One-BS network with active UE-VBSs represented by black nodes and UEs represented by white nodes [3]

ML algorithms (i.e., clustering, classification); and iii) finally, the last part, utilisation of classification variable shows how to use the trained result of the machine learning algorithm after training for the prediction and classification.

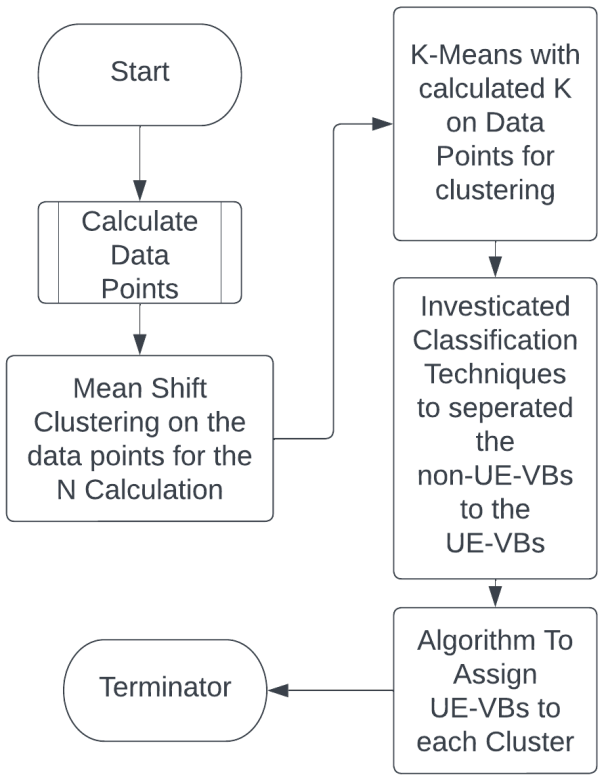

Fig. 3. Flowchart of Implementation

A. Data Selection

This section shows how the data are generated and how the clustering is executed in order to use the classification methods. Moreover, it shows the data engineering related to the simulating attributes of the user equipment.

To generate data, MATLAB software was initially used to simulate two attributes of the user equipment data: the distance of the user equipment from the base station and the discharge rate of the user equipment. Fig. 4 and 5 show the entire UE layout around a BS. The distance from the base station was calculated using the Euclidean distance formula, and the battery discharge rate was selected from 25m to 750m and 1000 mAh to 5000 mAh, respectively. This was because the discharge rate depended on how the users used their devices and the workload on the device. Fig. 4 shows the placement of 100 UEs. Following these two attributes, the data was simulated for various other characteristics/attributes of the mobile device by utilising Python and, more specifically, by using the method sample of Python (One million values were sampled using Python's standard sample() function) for the training of the investigated classification ML approaches.

Thus, for the first stage of our investigation, which is clustering, we utilise Mean Shift Clustering with the calculated points. We use the Cartesian distance as a metric to calculate the K number. The K number is the best number of clusters we can have under the Base Station. Then, we use K as the number of clusters in the K-Means and the calculated points to separate them into clusters. Next, in the second stage, the classification ML techniques that are trained with the aforementioned data are used to identify the UEs that can be UE-VB. In our examination, we select the best classification technique according to accuracy for the aforementioned task.

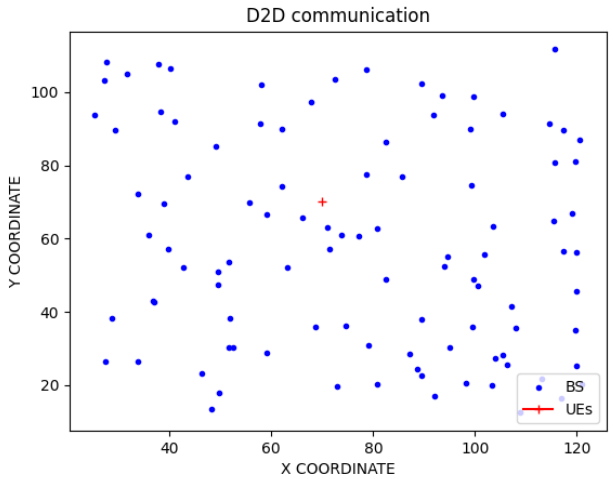

Fig. 4. Position of 100 UE's

B. Features/Metrics Used to Train the ML approaches

This section presents the features/metrics used to train our investigated approaches. The features/metrics are the follow-

ing:

- Data Rate: The data rate is in the range of 1 Mbps to 5 Gbps for the current generation of mobile devices, and hence, we assume that the basic range of data rates is about 1 Mbps to 5 Gbps and 1 million values were sampled from this range.
- SNR: From the literature survey, it is observed that the range of SNR lies between 0.014 and 1000. Therefore, the sample values are multiplied by 1000, and the final column is divided by 1000 to get the old expected values.
- Quality of Service or QoS: It is assumed that the Quality of Service exists between 0.2 and 0.99. However, for sampling values, values are used in percentages, i.e., 20% and 99% are used.
- Battery Capacity: Multiple models of the current generation of smartphones were reviewed, and the expected battery capacity range was observed and employed. For this experiment, battery capacities were taken from 1000 mAh to 5000 mAh with 250 mAh intervals.
- Storage: The current generation of mobile devices allows for modular and expandable storage; hence corroborating with current mobile devices is futile. Instead, values from the following list were sampled: [32, 64, 128, 256, 512], which was consistent with the current generation of mobile devices.
- Battery Level: This attribute considers the percentage level of battery left in the user equipment. Therefore, values were sampled from 1 to 100 to create types of equipment with different battery levels left on the device.
- CPU name/model, speedof CPU and RAM: Unlike battery capacity and storage, the CPU model, speed of the processor, and RAM are interdependent. A processor's speed depends on which series it belongs to, and mobile device RAM is selected not to bottleneck the devices' processors. This meant that it was necessary to research the market for processors that were and are used by the majority of the population and find corroborating values of RAM for mobile devices with the same processor. Finally, a data frame was created so the values would not change. This was done to ensure that a device with one type of processor would have its respective processor speed and RAM during sampling.

After simulating the data, a data frame with the attributes such as distance from the base station, discharge rate, data rate, SNR, QoS, battery capacity, processor name, and speed of the processor and RAM capacity, along with one million data points, is created.

C. Utilisation of Classification Variable

After creating the data attributes, the target classification variable was added for the machine learning model, which was whether a UE is VBS or non-VBS. Specifically, we have assigned one or 1 in the training part of the output value feature "Is_VBs?" if the UE achieves 60% or more for each feature/metric, depending on the highest value of each of the features/metrics mentioned in Section IV-B; otherwise, we have assigned zero or 0. On this basis, we created an algorithm that, following classification, uses the resulting classification score to select the most VB-capable UEs to become UE-VBs. This was accomplished by calculating a score for these attributes following the calculation of the categorical value of each UE from the ML approach and using the resulting score of classification value as a cutoff if the UE achieves less than the 60th percentile. Therefore, all the data points beyond this 60th percentile of the score were labelled as VBS or 1, and all values under the 60th percentile were marked as non-VBS or 0. The proposed algorithm that utilises the threshold and the score of the categorical values in order to select the UE-VBs of a cluster is shown in Algorithm 1.

Algorithm 1 Selection of UE-VBs

```
procedure (selection_of_UEVBs) cluster, Cluster_UEs
    BS                                  ▷ The Base Station
    dt_score_evaluator                  ▷ the Decision Tree that returns score value for a specific UE
    VBs_threshold ← 60%
    VBs_ListScores                      ▷ A Max Heap Binary Tree data structure that holds the VBs and in the first position the maximum value UE
    while Cluster_UEs is not empty do   ▷ read all UEs from the set
        Remove a UE from Cluster_UEs
        UEScore ← dt_score_evaluator(UE)   ▷ calculate the score of dt
        if UEScore ≥ VBs_threshold then
            Add UE with the UEScore in the VBs_ListScores
    if ∃ UEmax ∈ VBs_ListScores then
        UE_VBs(cluster) ← UEmax
    else
        UE_VBs(cluster) ← BS
```

Note that 1 million UEs/points were generated, 400,000 of these data points were labelled as VBS, and 600,000 were labelled as non-VBS. Though there is a class imbalance, this was done intentionally to accommodate the fact that there is a higher percentage of user equipment that would not act as VBS than the ones that would.

V. CLUSTERING

This section provides the clustering analysis done with the K-Means in conjunction with the means shift approach. Clustering (also known as cluster analysis) is the process of grouping a set of objects so that objects in the same group (called a cluster) have more comparable characteristics than those in other groups (clusters). Organising all UEs into clusters based on one or more comparison qualities (metrics) is essential. In our examination, we used the K-Means algorithm, the distance between points for clustering, and the mean shift clustering to calculate the number of clusters. The clusters

TABLE I
COMPARISON OF VARIOUS ML MODELS

Method	Training Accuracy (%)	Test Accuracy (%)	Error Rate (%)	Confusion Matrix
Decision Tree	100	99.9912	0.0001	[[150058 14] [8 99920]]
Logistic Regression	89.664	89.61	0.10386	[[13364 16425] [9541 90387]]
Linear Discriminant Analysis	89.664	96.893	0.0310679	[[142305 7767] [0 99928]]
Quadratic Discriminant Analysis	89.66	99.929	0.0007	[[149988 84] [91 99837]]
Linear SVC	89.66	99.97	0.000228	[[150035 37] [20 99908]]

produced by the clustering algorithm are depicted in Fig. 5, and the result of the Mean Shift Clustering approach for the K value was 4 or four.

Fig. 5 shows the clustering results after obtaining the cluster size using mean shift clustering, which was four, and the clustering is done using K-means clustering.

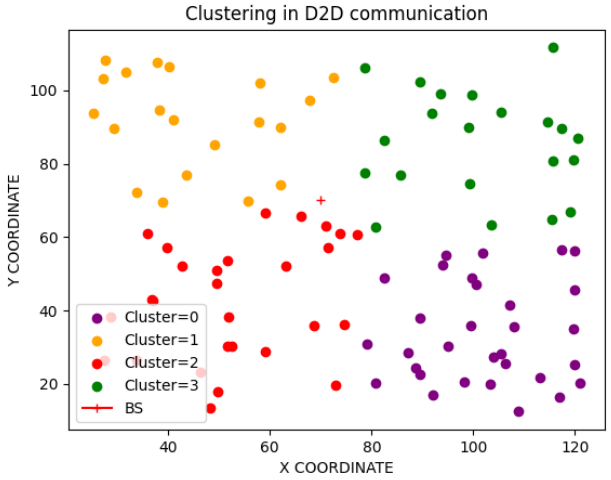

Fig. 5. Layout of experimental setup

In order to cross-check the K value returned by the Mean Shift Clustering approach, we utilised the Elbow method. The optimal number of clusters (K), the only parameter examined for the K-means clustering algorithm, is of the utmost importance. This is determinable using the Elbow approach. Using the elbow technique, K-means is executed multiple times with the k value incremented after each iteration, and the Sum of Squared Errors (SSE) is recorded. This decrease in SSE value at a given cluster size is known as the elbow [18].

From Figure 6, we can see that the SSE keeps decreasing as the number of clusters increases (k). The point where the SSE curve is called the elbow point, and the value of k at this point can be used as the best cluster size.

VI. CLASSIFICATION

Classification is a technique implemented by supervised learning, and within the classification, a class under which an object will fall is predicted. Concerning the problem statement, the target is a binary categorical variable with a value of zero/0 for devices not selected as VBS and one/1 for devices that can act as VBS. The data are split into training and testing, with

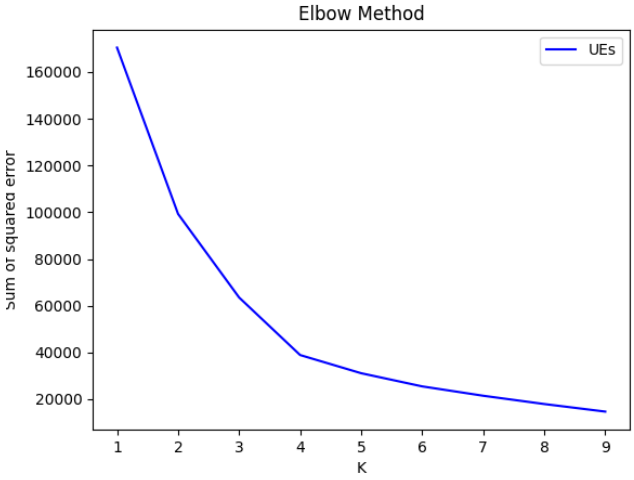

Fig. 6. Sum of Squared Errors (SSE) vs Number of clusters

80% used for training and 20% for testing. Table I shows the machine learning algorithms used for prediction with the calculation of the accuracy and error rate from the confusion matrix.

As shown in Table I, the decision trees algorithm outperforms the other classification algorithms. This can be attributed to the problem being a binary classification problem, and decision trees, an ensemble method, performs and scales well in cases with binary target variables. Therefore, the decision trees algorithm was chosen to be the best of all the algorithms.

In Figure 7, we demonstrate the resulting UE-VBS after running the proposed algorithm shown in Section IV-C on the clustered data shown in Section IV-A along with the use of Decision Tree for UE-VBs selection of each cluster.

VII. CONCLUSION AND FUTURE WORK

In order to enable a network area to relieve the existing traffic in the network with the use of machine-learning techniques. A network was simulated and split into clusters with the use of clustering, and a single UE was selected with the aid of classification to act as a VBS for each identified cluster. In this research, we will investigate the joined clusterisation and classification of UEs to UE-VBs problem in D2D communications using different competing ML techniques for classification in addition to K-Means, Means-Shift, and decision trees for clusterisation and classification, respectively. K-Means, in conjunction with Mean Shift Clustering for cluster number

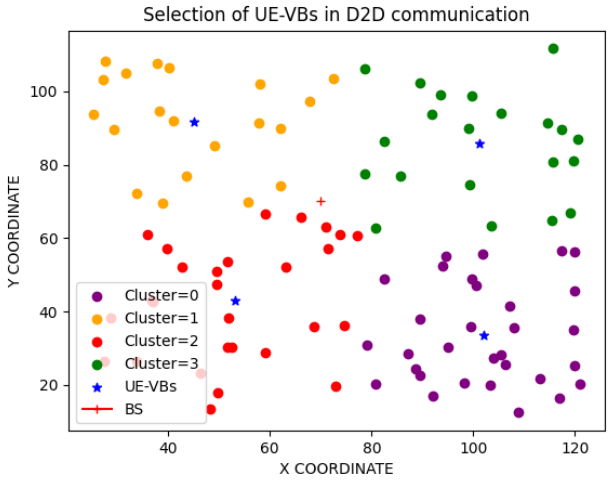

Fig. 7. Layout of experimental setup

determination, aims to cluster and group the simulated network devices into areas through the construction of clusters. In our simulation, 100 devices were simulated and then clustered. For each cluster, data were constructed (all 11 features but not the target variable), and then the saved machine learning model was invoked for selecting VBS and non-VBS UEs. After selecting candidate UEs for VBS for each cluster with the use of the decision trees, which was the most accurate among the rest of the investigated classification ML approaches, a score was formulated based on the features of UEs. Furthermore, the one with the maximum score was selected as UE-VBS. After choosing the clusters based on features, the cluster head also determined which possessed the best score. In future work, we will investigate how the selection of the UE-VBs over non-UE-VBs helps the D2D communication network to be stable, providing long-term stability and security, better spectral efficiency, and less power consumption.

ACKNOWLEDGEMENT

This research is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº739578 and the government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development. This work has received funding from the European Union under grant agreement No 101087257. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This work has received funding from the European Union under grant agreement No 101095363. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] I. Ioannou, V. Vassiliou, C. Christophorou, and A. Pitsillides, "Distributed Artificial Intelligence Solution for D2D Communication in 5G Networks," *IEEE Systems Journal*, vol. 14, pp. 4232–4241, sep 2020.

[2] I. Ioannou, C. Christophorou, V. Vassiliou, and A. Pitsillides, "A novel distributed ai framework with ml for d2d communication in 5g/6g networks," *Computer Networks*, vol. 211, 2022.

[3] P. Swain, C. Christophorou, U. Bhattacharjee, C. M. Silva, and A. Pitsillides, "Selection of ue-based virtual small cell base stations using affinity propagation clustering," *in the International Wireless Communications Mobile Computing Conference (IWCMC)*, vol. 2018, pp. 1104–1109, June 2018.

[4] X. W. al., "Green virtual base station in optical-access-enabled cloud-ran," in *2015 IEEE International Conference on Communications (ICC)*, pp. 5002–5006, June 2015.

[5] Z. Z. al., "Virtual base station pool: towards a wireless network cloud for radio access networks," in *Proceedings of the 8th ACM International Conference on Computing Frontiers*, (NY, USA), pp. 1–10, New York, May 2011.

[6] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyperdense heterogeneous and small cell networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 20–27, 2013.

[7] C. Christophorou, A. Pitsillides, and I. Akyildiz, "Celec framework for reconfigurable small cells as part of 5g ultra-dense networks," *IEEE International Conference on IEEE in Communications (ICC)*, vol. 2017, pp. 1–7, 2017.

[8] C. B. Lucasius, A. D. Dane, and G. Kateman, "On k-medoid clustering of large data sets with the aid of a genetic algorithm: background, feasibility and comparison," *Analytica Chimica Acta*, vol. 282, no. 3, pp. 647–669, 1993.

[9] L. Gavrilovska, V. Rakovic, and V. Atanasovski, "Visions towards 5g: Technical requirements and potential enablers," *Wireless Personal Communications*, vol. 87, no. 3, pp. 731–757, 2016.

[10] T. Pfeiffer, "Next generation mobile fronthaul architectures," in *Proc. OFC*, (US), Los Angeles, 2015.

[11] Z. K. al., "ebase: A baseband unit cluster testbed to improve energy-efficiency for cloud radio access network," in *Proc. ICC*, Budapest, 2013.

[12] K. P. S. and, "and m. -s," *Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access*, vol. 8, pp. 80716–80727, 2020.

[13] K.-L. Wu and M.-S. Yang, "Mean shift-based clustering," *Pattern Recognition*, vol. 40, no. 11, pp. 3035–3052, 2007.

[14] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *J. Chemometrics*, vol. 18, pp. 275–285, 2004.

[15] F. Thabtah, N. Abdelhamid, and D. A. Peebles, "machine learning autism classification based on logistic regression analysis," *Health Inf Sci Syst*, vol. 7, p. 12, 2019.

[16] B. Ghojogh and M. Crowley, "Linear and quadratic discriminant analysis: Tutorial," preprint, arXiv, 2019.

[17] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and non-linear pattern classification," in *2019 International Conference on Intelligent Sustainable Systems (ICISS*, pp. 24–28, 2019.

[18] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," *IOP Conference Series: Materials Science and Engineering*, vol. 336, p. 012017, apr 2018.