A Comparative Study on Machine Learning Algorithms for Indoor Positioning

Sinem Bozkurt[1], Gulin Elibol[2]
Dept. of Computer Engineering[1]
Dept. of Electrical & Electronics Engineering[2]
Eskisehir Osmangazi University
Eskisehir, Turkiye
sbozkurt@ogu.edu.tr
gelibol@ogu.edu.tr

Serkan Gunal
Dept. of Computer Engineering
Anadolu University
Eskisehir, Turkiye
serkangunal@anadolu.edu.tr

Ugur Yayan
R&D Department
Inovasyon Muhendislik Ltd. Sti.
Eskisehir, Turkiye
ugur.yayan@inovasyonmuhendislik.com

***Abstract*—Fingerprinting based positioning is commonly used for indoor positioning. In this method, initially a radio map is created using Received Signal Strength (RSS) values that are measured from predefined reference points. During the positioning, the best match between the observed RSS values and existing RSS values in the radio map is established as the predicted position. In the positioning literature, machine learning algorithms have widespread usage in estimating positions. One of the main problems in indoor positioning systems is to find out appropriate machine learning algorithm. In this paper, selected machine learning algorithms are compared in terms of positioning accuracy and computation time. In the experiments, UJIIndoorLoc indoor positioning database is used. Experimental results reveal that k-Nearest Neighbor (k-NN) algorithm is the most suitable one during the positioning. Additionally, ensemble algorithms such as AdaBoost and Bagging are applied to improve the decision tree classifier performance nearly same as k-NN that is resulted as the best classifier for indoor positioning.**

Keywords—indoor positioning; Received Signal Strength (RSS); classification; machine learning algorithms; nearest neighbor (NN); SMO; decision tree (J48); Naïve Bayes; Bayes Net; AdaBoost; Bagging; WEKA; RF Map; Localization.

I. Introduction

Indoor positioning systems are becoming widespread. Since GPS is available only at outdoor locations, indoor positioning systems have applications such as mobile robots localization [1] and source localization [2] and used to estimate the location of a subject in indoor areas like hospitals, libraries, airports, malls or warehouses.

The fingerprinting based positioning is a commonly used one consisting of two phases that are online (training) and offline (positioning). In offline phase, the radio map is created via measuring RSS values from existing access points in environment. Radio maps include not only RSS values but also coordinate information of the points where the measurements were taken from [3]. In addition to this, the floor number, type of mobile unit etc. may also be stored in the map [4]. In online phase, localization is accomplished by matching the RSS values of the radio map and the RSS values measured by the mobile unit [5]. There are some problems in fingerprinting methods due to radio maps and the algorithms used in positioning. The types of mobile devices that are used for creating radio maps in training phase can be different from the ones that are used during positioning. Besides, the number of access points in environment or places of them can be changed. These problems negatively affect the accuracy of positioning. The selection of algorithms and parameters is another factor affecting the accuracy of positioning in online phase. In this part, machine learning algorithms are used to estimate the position. In applications, predicting symbolic locations (offices, labs etc.) instead of physical coordinates is treated as classification problem [6]. In [7], different machine learning algorithms are compared in terms of accuracy and processing time to determine the most suitable algorithm in indoor positioning.

In this study, an extensive analysis is carried out to determine appropriate classification algorithm to resolve indoor positioning problems. During the experiments, UJIIndoorLoc database, which is prepared for indoor positioning systems [8], is used. The classification is performed firstly using original dataset considering RSS values from 520 wireless access points (WAPs) and newly defined attribute named as “cell” that compose the attributes BuildingID, Floor, SpaceID and RelativePos from the original data set. Then, a new method is proposed named as “Deductive Separation for Indoor Positioning (DESIP)”. In this method, first of all, only the building information and RSS values that are measured from 520 WAPs are used for the classification task. In the first classification process, the building information is defined as the class. Later, the database is divided into 3 parts on the basis of building information and is classified according to the floor information. Predicting the building of each test data is done in the first classification and the floor of the test data, whose building information is found in the first classification, is predicted in the second classification stage. Finally, every dataset, which is already split based on building information in the second classification, is split again on the basis of floor

978-1-4673-7751-5/15/$31.00 ©2015 IEEE

information. Each dataset divided based on the floor information are classified using space and relative position features. In the experiments, deterministic algorithms such as nearest neighbor (NN), SMO, decision tree (J48) and probabilistic algorithms such as Naïve Bayes and Bayes Net are employed. The most appropriate algorithm for the solution of indoor positioning problem is determined by comparing the accuracy and computational time of each approach. In addition to these tests, ensemble learning algorithms, namely AdaBoostM1 and Bagging, are used to improve the performance of the selected classifier J48. J48 is chosen because its performance is enhanced nearly the same as NN in indoor positioning when these ensemble algorithms are run together with J48. WEKA [9] library is utilized during the experiments.

The rest of the paper is organized as follows: Section II presents the related work in the literature. Section III explains UJIIndoorLoc database used in the experiments. Section IV describes the classification algorithms whereas Section V presents the experimental work. Finally, conclusions and future work are given in Section VI.

II. RELATED WORK

Position estimation methods are categorized as deterministic and probabilistic. In deterministic methods, environment is divided into cells to establish a radio map and estimated position is obtained by finding the best match between the new measurement and the measurements in the radio map [10, 11, 12, 13]. In probabilistic methods, also called as distributed-based methods, signal strength distributions obtained from Access Points (APs) are used to construct radio map and probability distribution functions are used to estimate the position of subject [14, 15, 16, 17, 18].

In [10], nearest neighbor algorithm is used to find the floor in tall and multi floor building using Euclidean distance metric. K-nearest neighbor is applied in [11], where radio map is constructed to use in comparison of each signal strength value in online phase and mean of the best matched K position is determined as the estimated position. In [12], an indoor localization method is proposed for Zigbee sensor networks using classification of link quality patterns between each reference node and a target node in a specific location rather than using calculation with signal strength, arrival time, or angle. To classify link quality patterns for each location, an artificial neural network (ANN) classifier is used. APs and fingerprints filtering approaches are proposed for filtering the WiFi positioning radio maps to reduce the computation overhead and increase positioning accuracy in [13]. Unused access points and worse fingerprint samples are removed using classification rules from radio map in APs filtering and fingerprints filtering approach, respectively. In this study, J48 decision tree algorithm is used as the classification algorithm.

Particle filter, a Bayesian based method, is employed in [14]. Observed signal strengths are obtained using Bayesian inference in [15] and the estimated position is determined as the highest probability in the resulting distribution. An extended Kalman filter based approach is presented in [16], where the intra cell position of a cellular device is estimated using RSS readings from base stations. This estimate, movement pattern data and velocity vectors are combined in order to predict the next cell crossing. In [17], a Bayesian filter based approach is proposed. In this study, a posterior probability distribution over the target's location is obtained by inverting Bayesian belief network. In [18], subset of the strongest access points is considered instead of all access points and target location is predicted using Bayesian estimate.

III. INDOOR POSITIONING DATASET

UJIIndoorLoc, which can be downloaded from UCI Machine Learning Repository, is the most comprehensive indoor positioning database in the literature. This database includes 3 buildings that have 4, 4 and 5 floors respectively of Universitat Jaume I that is located on the land of $108.703m^2$. While creating the radio map, 25 different kinds of mobile units are used to take measurement from 933 different reference points by 20 different users. The entire database is separated such that 19.937 records are reserved for training and 1.111 records are reserved for testing. There are 529 features and these features are the coordinates where WiFi fingerprints are taken, such as building, floor, space (office, lab, etc.), relative position (in a room or at corridor) etc. In this work, firstly original UJIIndoorLoc training dataset including RSS values from 520 WAPs and a new attribute "cell" that composes the attributes floor, buildingID, spaceID and relativeposition from the original data set is used for the classification task. The steps of the experiments using this dataset are shown in Fig. 1.

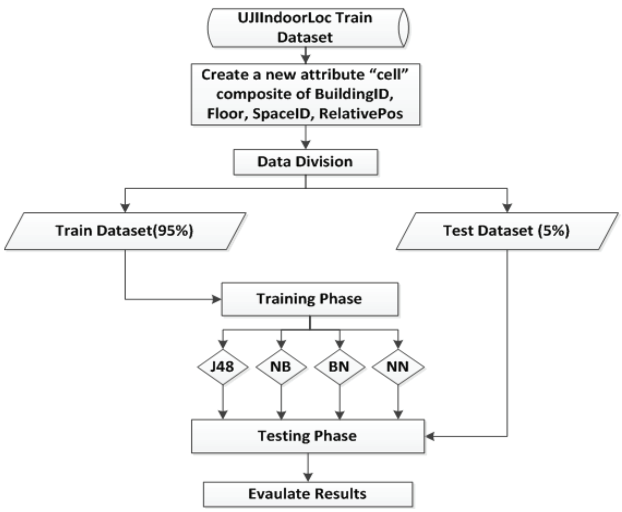

Fig. 1 The new attribute "cell" construction phase

Two new datasets are constituted from using only the training dataset for classification experiments. These new datasets are used in the training and testing phases of classification and named as IP_Train and IP_Test, respectively. The aim of the study is comparing the accuracy of the algorithms while classifying building, floor and region (space and relative position). IP_Train dataset is 95% and IP_Test dataset is 5% of all records.

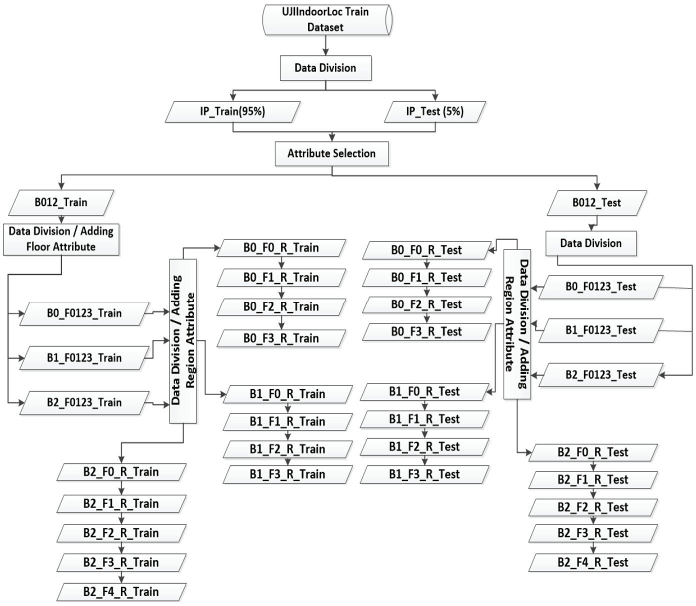

Fig. 2 Dataset constitution phase according to DESIP approach

The classification is performed in 3 phases, namely building, floor and region, according to DESIP approach that is shown in Fig. 2. The datasets that are used in these phases i.e. steps of the DESIP method are clearly defined below.

A. *Building Classification*

B012_Train and B012_Test datasets are created from IP_Train and IP_Test, respectively. Each of these datasets are arranged to include RSS values from 520 WAPs and all building information. The attributes of B012_Train and B012_Test datasets are *WAP001, WAP002 … WAP520, BuildingID.*

B. *Floor Classification*

IP_Train and IP_Test datasets are split into three for different buildings. B0_F0123_Train, B1_F0123_Train and B2_F01234_Train datasets are formed from IP_Train and B0_F012_Test, B1_F0123_Test and B2_F01234_Test datasets are formed from IP_Test. These six datasets are arranged to include only one building and all floor information of this building unlike the building classification datasets. The attributes of floor classification are *WAP001, WAP002, … WAP520, BuildingID, Floor.*

C. *Region Classification*

The datasets that are already split into floor classification are split again on the basis of region. For instance, datasets of B0_F0_R_Train, B0_F1_R_Train, B0_F2_R_Train and B0_F3_R_Train are created from B0_F0123 dataset. The same splitting operations are done for the rest of the training and test floor datasets. A new feature is defined for splitting on the basis of space and relative position. This feature is the combination of space and relative position information and named as "*region*". This feature is added to the datasets that are split for region classification. This feature is illustrated in Table 1. Features of SpaceID and RelativePosition are explained briefly in [4].

Table 1. Relation between Region and SpaceID, RelativePosition

SpaceID	RelativePosition	Region
100	1	200
100	2	201

The attributes of region classification are *WAP001, WAP002 … WAP520, BuildingID, Floor, Region*.

IV. Classification Algorithms

In the following section, the classification algorithms employed in this study are briefly described.

A. *Decision Tree*

Decision tree [19] is a well-known and commonly used method in machine learning. A decision tree is a hierarchical structure including decision (non-terminal) nodes, branches and leaf (terminal) nodes that represent attributes (features), conditions and classes, respectively. Entropy or information gain can be used to create nodes in a decision tree. Each non-terminal node that contains a condition is used to determine which branch to follow from that node. If the condition is true then algorithm will follow one branch otherwise it will follow the other one. When the algorithm reaches a leaf node, then the label stored in the leaf returned as a class. Quinlan's ID3 and its successor, C4.5, are the most popular ones among decision tree algorithms [19]. J48 decision tree in WEKA uses entropy and implements Quinlan's C4.5 algorithm to generate a pruned C4.5 tree. It first creates a decision tree to classify a new item using the attributes of training data. It then chooses the attribute that most obviously discriminates the various instances and looks for another attribute that gives the highest information gain. It continues the process until finding subset instances belonging to the same class, and so the leaf node is created, and it stops when it handles all of the attributes [20, 21].

B. *Naïve Bayes*

A Naïve Bayes [22] classifier based on Bayes Theorem is a supervised learning algorithm [23]. It is robust to noisy data, easy to build, shows high accuracy and speed when applied to large databases and performs more complicated classification models. Hence, it is widely used in classification tasks. It calculates probability of each attribute in the data assuming that they are equally important and independent of each other.

This assumption is called as class conditional independence [24, 25].

C. Bayesian Network

The Bayesian Network algorithm is widely used for classification and is based on the Bayes theorem where conditional probability on each node is calculated and forms a Bayesian Network. It is also called Belief Network or Casual Network. Bayesian Network has two parts named as qualitative and quantitative, which are topological structure of Bayesian Network and conditional probability table (CPT), respectively [26]. The network is constructed using several parameters, including estimator and search method. Bayesian Network is a directed acyclic graph where each node represents an attribute of the data and a set of probability distributions. These distributions give the probabilities for the value of each node given that parents of node. In the case of discrete variables, the probabilities are encapsulated by a CPT for each node, which enumerates all possible combinations of values of the parent attributes of node and the corresponding conditional probability for each value of the node.

D. K-Nearest Neighbor

K-Nearest Neighbor (K-NN) [27] classifier is also known as a distance based classifier that classifies instances based on their similarity. It is one of the most popular algorithms in machine learning and is a type of Lazy learning in which the function is only approximated locally and all computation is delayed until classification. The unknown tuple in K-NN is assigned to most common class among its K-nearest neighbors. When K = 1, the unknown tuple is assigned the class of the training tuple that is closest in the pattern space [28].

E. SMO

Sequential minimal optimization (SMO) [29] algorithm is represented by John C. Platt for training the support vector classifier using polynomial or RBF kernels. It is one of the most common algorithms for large-margin classification by SVM. It globally replaces all missing values and transforms nominal attributes into binary ones. SVM is a classification technique based on neural network technology using statistical learning theory [30]. It looks for a linear optimal hyper plane so that the margin of separation between the positive class and the negative class is maximized. In practice, most data are not linearly separable; thus, to make the separation feasible, transformation is performed using a Kernel function. The input is transformed into a higher dimensional feature space using nonlinear mapping [30]. A decision on the Kernel function is needed in implementing SVM. The kernel defines the function class [31].

F. AdaBoost

AdaBoost (Adaptive Boosting) [32] is an ensemble learning algorithm. Generally, it can be used with weak machine learning algorithms to improve their performance. It is simple to implement, fast, and less susceptible to overfitting. It improves unstable classification algorithms such as J48, DecisionStump, etc. The idea behind this algorithm is to obtain highly accurate classifier by combining many weak classifiers. It works by repeatedly running a given weak learning algorithm on various distributions over the training data, and then combining the classifiers produced by the weak learner into a single composite classifier [33]. The classifiers in the ensemble are added one at a time so that each subsequent classifier is trained on data which have been difficult for the previous ensemble members. Weights are set to instances in the data set, in a rule that the instances that are difficult to classify get more weight. This rule drives subsequent classifiers to focus on them [34].

G. Bagging

Bagging [35] builds bags of data of the same size of the original data set by applying randomly selecting different subsets of the training data with many examples appearing multiple times. This process named as bootstrap replicate of the training data. The idea behind this technique is to build various classifiers by using these subsets. Each subset is used to train one individual classifier. This ensemble approach uses number of classifier as a priori [35].

V. Experimental Work

Experiments were performed in WEKA using UJIIndoor original training dataset and the datasets that are constructed according to DESIP method. Our goal was to evaluate performance of the classification algorithms for indoor positioning. For this purpose, NN, SMO, J48, Naïve Bayes and BayesNet algorithms were comparatively tested. The flowchart given in Fig. 3 shows the implementation of these algorithms to the datasets defined in Fig. 2.

Accuracy and computation time were used as the performance metrics in comparison of the classifiers. True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) measures were employed to obtain accuracy. These measures are acquired from confusion matrix that calculates the number of instances predicted correctly or incorrectly by a classification model. The possibility that the classifier can correctly predict positive and negative instances named as accuracy and is calculated using Eq. 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

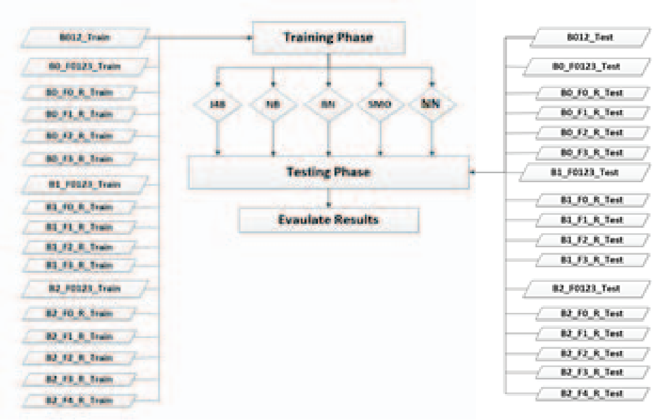

Fig. 3 Training and Testing Phase

In the experiments, PolyKernel was used for SMO as a Kernel function. In BayesNet algorithm, SimpleEstimator method and K2 were chosen as the estimator and search algorithm. They are optimal parameters for WEKA. Firstly, we applied UJIIndoorLoc whole training dataset using RSS values obtained from 520 WAPs and new attribute “cell” that includes BuildingID, Floor, SpaceID and RelativePos information. J48, NB, BN, NN and SMO algorithms were used. In results, J48 gives the best accuracy (99.89%) and BN gives the worst accuracy (0.42%). Since SMO could not complete the classification task, its accuracy results were not given. The accuracy results of this experiment are shown in Fig. 4.

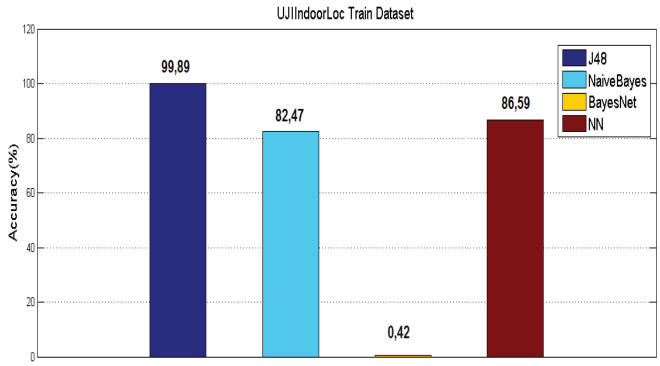

Fig. 4 Accuracy results of classification using whole dataset

The first step of indoor positioning classification using DESIP method is building classification as mentioned in Section III. The accuracy results of the algorithms for this are shown in Fig. 5. It is clear from this figure that BayesNet gives 99.80% accuracy which is better than that of other classifiers.

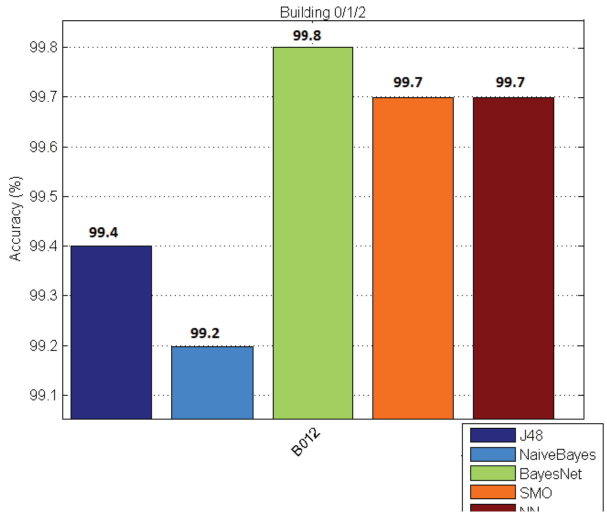

Fig. 5 Accuracy results of building classification

In the second step, the classification was performed based on the floor attribute in each building. The results are provided in Fig. 6. One can note from this figure that NN has the best results (averagely 98.5), J48 is the runner up, and SMO takes the third place.

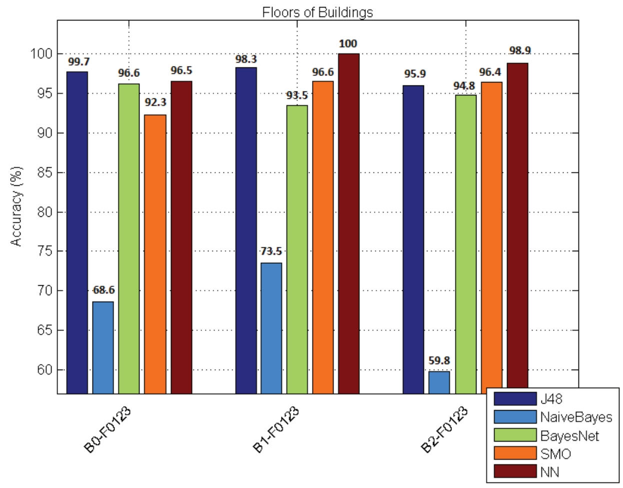

Fig. 6 Accuracy results of floor classification

Classification was performed using new attribute, region, in the last step. The results are given in Fig. 7. NN offers the best results again and SMO becomes the runner up.

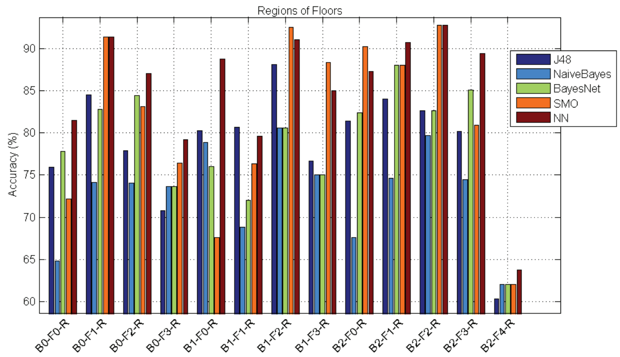

Fig. 7 Accuracy results of region classification

From Fig. 8 and Fig. 9, we can consider that NN is the superior algorithm with respect to other methods. It provides higher accuracy and requires less execution time. Comparison of accuracy and time values are illustrated in Table 2 and 3, respectively.

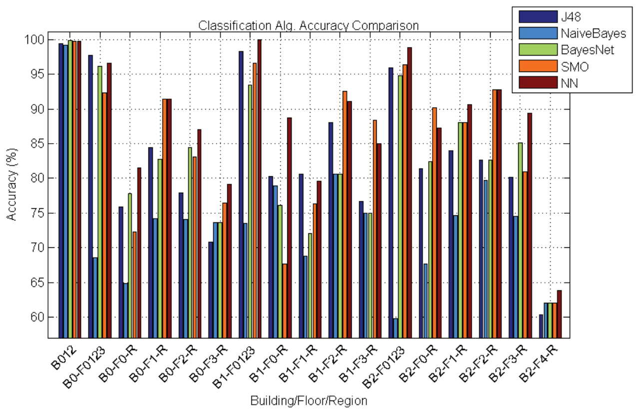

Fig. 8 Accuracy results of all classifications

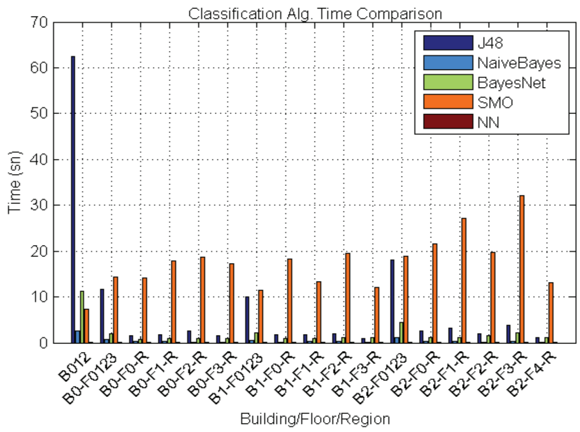

Fig. 9 Computational time of classification methods

Table 2. Accuracy results of machine learning algorithms

	J48	Naive Bayes	Bayes Net	SMO	NN
B012	99,4	99,19	99,80	99,69	99,69
B0_F0123	97,7	68,58	96,17	92,34	96,55
B0_F0_R	75,93	64,81	77,78	72,22	81,48
B0_F1_R	84,48	74,14	82,76	91,37	91,38
B0_F2_R	77,92	74,03	84,42	83,12	87,01
B0_F3_R	70,83	73,61	73,61	76,39	79,17
B1_F0123	98,28	73,54	93,47	96,56	100
B1_F0_R	80,28	78,87	76,06	67,61	88,73
B1_F1_R	80,64	68,82	72,04	76,34	79,57
B1_F2_R	88,06	80,60	80,59	92,54	91,04
B1_F3_R	76,67	75	75	88,33	85
B2_F01234	95,95	59,77	94,83	96,40	98,87
B2_F0_R	81,37	67,65	82,35	90,19	87,25
B2_F1_R	84	74,67	88	88	90,67
B2_F2_R	82,61	79,71	82,61	92,75	92,75

According to accuracy results in Table 2, Bayes Net provides the best accuracy (99.80 %) for the classification of building label. In all other cases, NN offers the highest accuracy values. The computational time of each classifier for each dataset are given in Table 3. Here, NN algorithm is again superior to all other methods for each data set.

Table 3. Elapsed time results of machine learning algorithms

	J48	Naive Bayes	Bayes Net	SMO	NN
B012	62,41	2,5	11,19	7,25	0,06
B0_F0123	11,74	0,65	2,03	14,31	0,01
B0_F0_R	1,5	0,22	0,7	14,1	0
B0_F1_R	1,71	0,21	0,88	17,93	0
B0_F2_R	2,52	0,2	0,9	18,77	0
B0_F3_R	1,62	0,2	0,88	17,3	0
B1_F0123	10,05	0,57	2,13	11,5	0,02
B1_F0_R	1,68	0,18	0,87	18,22	0
B1_F1_R	1,75	0,21	1,01	13,3	0
B1_F2_R	1,96	0,27	1,11	19,47	0
B1_F3_R	0,95	0,13	1,11	12,1	0
B2_F01234	18,09	1,18	4,4	18,81	0,03
B2_F0_R	2,54	0,24	1,11	21,57	0
B2_F1_R	3,27	0,29	1,15	27,1	0
B2_F2_R	1,86	0,2	1,46	19,76	0

As seen in the Table 2, J48 algorithm accuracy results are worse than NN. In order to approximate accuracy results of these two algorithms, we applied Adaboost1 and Bagging algorithms with J48. The results of this comparison are given in Fig. 10 as well.

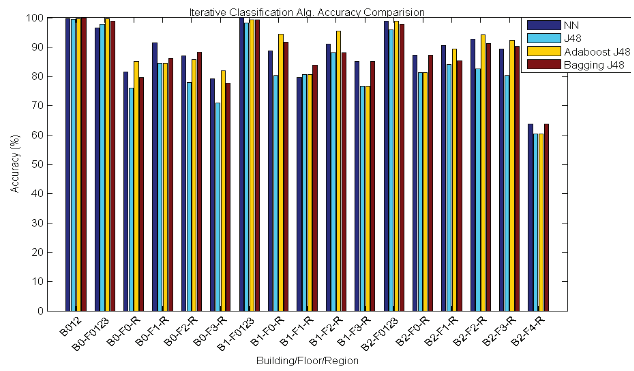

Fig. 10 Accuracy results of NN, J48, Adaboost with J48 and Bagging with J48

VI. CONCLUSIONS

In this study, selected machine learning algorithms are compared in terms of accuracy and computation time using UJIIndoorLoc database. Our aim is to find the most appropriate classifier for indoor positioning problem. As a conclusion, we deduced that NN is superior to all other methods to estimate position. Besides, J48 provides nearly same performance when used with iterative algorithms, namely AdaBoost and Bagging. Selection of discriminating features and using different performance metrics will be interesting and useful future works. Additionally, we are planning to constitute our own dataset considering additional features to improve performance of the indoor positioning system.

ACKNOWLEDGMENT

This work is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under grant number 1130024.

REFERENCES

[1] U. Yayan, H. Yücel, and A. Yazıcı, "A low cost ultrasonic based positioning system for the indoor navigation of mobile robots", Journal of Intelligent & Robotic Systems, vol. 78, no. 3, pp 541-552, 2015.

[2] C. Yeniceri, T. Tuna, U. Yayan, A. Yazıcı, H. Yücel, and V. Bayar, "A smart solution for transmitter localization", International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2013), pp. 1-5, 2013.

[3] C. Laoudias, R. Piche, and C. Panayiotou, "Device self-calibration in location systems using signal strength histograms", Journal of Location Based Services, vol. 7, no. 3, pp. 165-181, 2013.

[4] J. Torres-Sospedra, R. Montoliu, A. Martınez-Usó, T. J., J. P., Avariento, M., Benedito-Bordonau, and J., Huerta, "UJIIndoorLoc: A New Multi-building and Multi-floor Database for WLAN Fingerprint-based Indoor Localization Problems", 5th International Conference on Indoor Positioning and Indoor Navigation, 2014.

[5] Y. C. Cehng, Y. Chawathe, A. LaMarca, and J. Krumm, "Accuracy characterization for metropolitan-scale Wi-Fi localization", 3rd International Conference on Mobile Systems, Applications, and Services, pp. 233-245, 2005.

[6] F. Seco, A. R. Jimenez, C. Prieto, J. Roa, and K. Koutsou, "A survey of mathematical methods for indoor localization", IEEE International Symposium on Intelligent Signal Processing (WISP 2009), pp. 9-14, 2009.

[7] S. Bozkurt, S. Günal, U. Yayan, and V. Bayar, "Classifier selection for RF based indoor positioning", Signal Processing and Communications Applications Conference (SIU), pp 791-794, 2015.

[8] http://archive.ics.uci.edu/ml/, UCI Machine Learning Repository.

[9] http://www.cs.waikato.ac.nz/ml/weka/, WEKA.

[10] A. Varshavsky, A. LaMarca, J. Hightower, and E. de Lara, "The SkyLoc floor localization system", Fifth Annual International Conference on Pervasive Computing and Communications (PerCom '07), IEEE, pp.125-134, March 2007.

[11] P. Bahl, and N. P. Venkata, "RADAR: An in-building RF-based user location and tracking system", Nineteenth Annual Joint Conference of IEEE Computer and Communications Societies (INFOCOM 2000), vol. 2, 2000.

[12] Y. Ha, E. Ae-cheoun, and B. Yung-cheol, "Efficient sensor localization for indoor environments using classification of link quality patterns", International Journal of Distributed Sensor Networks, 2013.

[13] S. Eisa, J. Peixoto, F. Meneses, and A. Moreira, "Removing useless APs and fingerprints from WiFi indoor positioning radio maps", International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp.1-7, Oct. 2013.

[14] V. Seshadri, V. G. Zaruba, and M. Huber, "A Bayesian sampling approach to in-door localization of wireless devices using received signal strength indication", Third IEEE International Conference on Pervasive Computing and Communications (PERCOM 2005), pp.75-84, March 2005.

[15] X. Chai, and Q. Yang, "Reducing the calibration effort for location estimation using unlabeled samples", Third IEEE International Conference on Pervasive Computing and Communications (PERCOM 2005), pp. 95-104, March 2005.

[16] T. Liu, P. Bahl, and I. Chlamtac, "A hierarchical position-prediction algorithm for efficient management of resources in cellular networks", Global Telecommunication Conference (GLOBECOM '97), IEEE, vol. 2, pp. 982-986, November 1997.

[17] M. Isard, and A. Blake, "Contour tracking by stochastic propagation of conditional density", Computer Vision (ECCV '96), pp. 343-356, 1996.

[18] M. A. Youssef, A. Agrawala, A. U. Shankar, and S. H. Noh, "A probabilistic clustering-based indoor location determination system", Tech. Report, University of Maryland at College Park, CS-TR 4350, March 2002.

[19] J. R. Quinlan, " *C4. 5: programs for machine learning*", Elsevier, 2014.

[20] P. Niken, and H. Ohwada. "Applicability of machine-learning techniques in predicting customer defection", Technology Management and Emerging Technologies (ISTMET), 2014 International Symposium on. IEEE, 2014.

[21] G. I. Salama, M. B. Abdelhalim, and A. Zeid, "Experimental comparison of classifiers for breast cancer diagnosis", IEEE 2012 7th International Conference on Computer Engineering & Systems (ICCES), 2012.

[22] G. H. John, and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers", 11th Conference on Uncertainty in Artificial Intelligence, pp., 338-345, 1995.

[23] C. Anuradha, and S. Dhall, "Software Defect Prediction Using Supervised Learning Algorithm and Unsupervised Learning Algorithm", 2013.

[24] W. Yotsawat, and A. Srivihok, "Inbound tourists segmentation with combined algorithms using K-Means and Decision Tree", 10th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp.189-194, 2013.

[25] S. Ureerat, and P. Singsri, "The classifier model for prediction quail gender after birth based on external factors of quail egg", IEEE 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2014.

[26] D. Yang, and L. Jin-lin, "Research on personal credit evaluation model based on bayesian network and association rules", 2007 International Conference on Wireless Communications, Networking and Mobile Computing, 2007.

[27] D. W. Aha, D. Kibler , and M. K. Albert, " *Instance-based learning algorithms*", Machine Learning, vol. 6, pp., 37-66, 1991.

[28] C. Shah, and A. G. Jivani, "Comparison of data mining classification algorithms for breast cancer prediction", 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp.1-4, 2013.

[29] J. Platt, "*Fast Training of Support Vector Machines using Sequential Minimal Optimization*", Advances in Kernel Methods - Support Vector Learning, 1998.

[30] P. Niken, and H. Ohwada, "Applicability of machine-learning techniques in predicting customer defection", IEEE 2014 International Symposium on Technology Management and Emerging Technologies (ISTMET), 2014.

[31] S. M. Obaidullah, K. Roy, and N. Das, "Comparison of different classifiers for script identification from handwritten document", 2013 IEEE International Conference on Signal Processing, Computing and Control (ISPCC), pp.1-6, 2013.

[32] Y. Freund, and R. E. Schapire, "Experiments with a new boosting algorithm", 3th International Conference on Machine Learning, San Francisco, pp. 148-156, 1996.

[33] R. Shams, and R. E. Mercer, "Classifying Spam Emails Using Text and Readability Features", 2013 IEEE 13th International Conference on Data Mining (ICDM, pp. 657-666, 2013.

[34] S. O. Sharif, L. I. Kuncheva, and S. P. Mansoor, "Classifying encryption algorithms using pattern recognition techniques", 2010 IEEE International Conference on Information Theory and Information Security (ICITIS), , pp. 1168-1172, 2010.

[35] L. Breiman, "Bagging predictors", Machine Learning. vol. 24, no. 2, pp.123-140, 1996.