Human activity recognition system using smartphone based on machine learning algorithms

Cite as: AIP Conference Proceedings **2555**, 040010 (2022); https://doi.org/10.1063/5.0110384
Published Online: 21 October 2022

Prakhar Gupta and Richa Arora

Human Activity Recognition System Using Smartphone based on Machine Learning algorithms

Prakhar Gupta[1,a)] and Richa Arora[1,b)]

[1]*CSE Department, Chandigarh University, Gharuan, Mohali, India*

Corresponding Author: [a)]*prakhargupta231@gmail.com,*
[b)]*richaaeri9988@gmail.com*

Abstract. Human Activity Recognition System aims to capture the state of the user with respect to the external/heterogeneous environment in order to analyze the human health conditions with the help of the numerous sensors attached to the different body parts of the human 3. The first hand-held mobile was developed in the year 1979 and from that year to 2011 it is surveyed that around 80% of the world population is now using the Smartphone and from that year it is been observed that there is been a tremendous increase/up gradation in the field of technology and size also and in this futuristic world of technology. As we can see now a days Smartphone are playing a key role and are used to collect the data of human activities and which is further used to monitor the health of the person. Prior to collect the human activity data we use to attach the sensors to the different body parts like chest, hands, thighs, wrist etc. In this paper, with the help of the m ach in e learn in g algorithms and the data mining techniques we are going to analyze the posture by keeping in mind about the attributes which are going to take part. We are using six of the attributes with the help of these attributes we are going to predict the conditions which are as follows: Sitting, Standing, Walking, Walking Upstairs, Walking downstairs and Laying.

Keywords—*Smart phone, activity recognition, machine learning, sensors, physical signals.*

I. INTRODUCTION

The first hand-held mobile was developed in the year 1979 and from that year to 2011 it is surveyed that around 80 % of the world population is now using the Smartphone and from that year it is been observed that there is been a tremendous increase/up gradation in the field of technology and size also and in this futuristic world of technology.[1] As we can see now a days Smartphone are playing a key role and are used to collect the data of human activities and which is further used to monitor the health of the person. Prior to collect the human activity data we use to attach the sensors to the different body parts like chest, hands, thighs, wrist etc. to check that how a human body performs with the change in external environment. But that was very difficult to wear those huge sensors and work in order the collect data.

But now with the upcoming of technology i.e. Smartphone, we can easily track for the human activity in order to collect the data.

Now these days' Smartphone is being equipped with the different sensors such as accelerometer, gyroscope etc. which help us to collect the data of human activities easily without having any problem. Now it's not required to wear those sensors in different body parts now you just need to do is carry your Smartphone with you and that is it [2].

In this novel also we are given with the dataset of the human activities which is further going to be used by us construct such algorithm which can give the accuracy as per our need i.e. to find the algorithm with the best fit accuracy by using the different machine learning algorithm [3].

II. PROBLEM PARAMETERS

In this proposed novel the problem parameters are related to the selection of the best fit algorithm which can be used to give the highest accuracy and also check with different software's which software are to be used for completing the whole process without any disturbance and can respond as per our needs [4].

Innovations in Computational and Computer Techniques
AIP Conf. Proc. 2555, 040010-1–040010-11; https://doi.org/10.1063/5.0110384
Published by AIP Publishing. 978-0-7354-4220-7/$30.00

There was a need for the dataset to make such model and for that we have taken the dataset from a known online repository which is UGC. It contains the dataset for the given problem dataset from UGC was used in spyder in order find out the best accuracy which can we get but these were not up to the mark. So we used another software which is weka and for that we have taken the dataset from the kaggle in .arff format in order to run the modal in the weka software to find out the best fit accuracy [5].

In this the problem was to find the right persons between the age group which we are required and on which we have to apply the algorithm of machine learning. As if we are not able to get the right person then it will be very difficult to find the accuracy as per our assumptions[6].

It was difficult to get the dataset for both the software's as it was not present with the weka library for the dataset for spyder it was easy to find the dataset. As dataset is the key role to find the result accurately.

III. PROPOSED PROBLEM

In the given novel the existing problem was solved using the different machine learning algorithm for which we need to check the accuracy for every algorithm and as much as can be fitted to find out the best fit accuracy for the proposed problem. There was a need to fit such algorithm which can give the accuracy as per our assumptions and to fulfill our assumptions there was a need to use not only spyder but also some other algorithm with some different software which can give the best results as per our need. There was a need for the dataset as well to make such model and for that we have taken the dataset from a known online repository which is UGC. It contains the dataset for the given problem dataset from UGC was used in spyder in order find out the best accuracy which can we get but these were not up to the mark. So we used weka software and for that we have taken the dataset from the kaggle in .arff format in order to run the modal in the weka software to find out the best f it accuracy.

IV. RELATED WORK

In this paper we have analyzed data set of different persons and use that data against various machine learning algorithms by using different platforms which help us to identify which methodology can give us the accurate results by applying different cross validation points across it. In general it is difficult to find algorithm which gives best accuracy for the given dataset a s there may be more than one algorithm having same accuracy [7].

We have used the experiences of the different researchers in order to write this paper. In this paper I have tried to find out the algorithms of the machine learning technique by which we can track down the activities of the humans in a new and creative way with the help of smartphones. As mentioned in the section there are three algorithms having same accuracy with ROC Curve having 10 Cross Validation and to find best of three we have decreased the value of cross validation to 5 and then to and then only we got required algorithm. Here in this novel we have learned as experiences of other researchers which was in the form of their research papers present on UCI repository [8]. And to know the concepts of the algorithm we used Wikipedia as a source of knowledge as well a s the research papers of those authors which used Weka as a tool to find the accuracy and also for introduction and abstract these papers helped me a lot. We used weka and spyder with tools that implement the algorithms and in those few are selected as they were giving better results than the other algorithms [9].

V. PROPOSED METHODOLOGIES

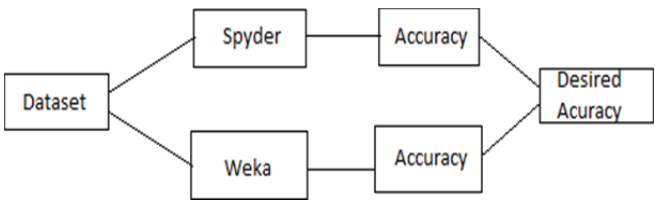

FIGURE 1. Methods used for activity recognition

In the above figure we have diagrammatically represented that how the operations are performed we have chosen the dataset then on that dataset we have performed the algorithms both from weka and spyder then the results that we got was in the form of accuracy. Here we have used different algorithms of weka as well as spyder but in weka we used a technique of cross validation and in that we have divided the whole dataset into groups and got the accuracy then we compared the accuracy of the algorithms from both software's and compared them to get the required accuracy[10]. We have implemented the cross validation in weka in order to get the accurate results even after getting the different sets of data like if the given data is in groups of 1, 2, 3… and with different groups accuracy may vary with respect to the different algorithms and in same algorithm it may vary with the use of different groups then here we find the algorithm whose value doesn't change with change in value of cross validation [11]. Whereas on spyder it was not that easy to divide the data into groups and after dividing we were not able to get the accuracy as per our needs so as per our study it was clear that the weka having algorithm are giving much better accuracies than the spyder. We have used below mentioned methods to determine the status of human activity to check the accuracy for the different activities performed by humans. These algorithms are, such as:-

A. i) Naive Bayes Classifier

It is also known as generative learning model. It is based upon the Bayes Theorem. In this classification the presence of any feature of dataset is unrelated to the feature of other type5. And if all these features are related to each other than to all these properties contribute to the probability of occurrence only. It is good make use of this classifier for large dataset and it is easy to work with[12].

B. ii) Logistic Regression

$$L = \beta 0 + \beta 2x2 + \beta 3x3$$

It is a supervised machine learning algorithm. It defines the relationship between the one dependent binary variable and one or more nominal. It is a statistical data set for analyzing a dataset in which are more independent variables that determine an outcome. The outcome which is observed is most probably only two outcomes not more than two. There should not exceed the possibility of more than two outcomes. The goal of this algorithm is to find the relationship between dataset and the two outcomes only[10].

C. iii) Decision Tree

$$P(A/B) = P(B/A)P(A) / P(B)$$

In this algorithm it finds the solution in the form of tree not in the form of data or in the form of table. In this dataset it breaks the data into the smaller and smaller subunits until it breaks into the smallest subunits. The final tree after the execution which is developed includes the decision and leaf nodes. It is one way to display an algorithm that only contains conditional control statements [13].

$$Gain\,(T, X) = Entropy\,(T) - Entropy\,(T, X)$$

D. iv) Nearest neighbor

In this it takes the labeled bunch of data to learn how to characterize the unlabeled data. To label any unlabeled data it takes the help of labeled data in order to characterize the labeled data. It is the proximity finding algorithm. In this we have to find the closet in the given dataset. It is based upon the optimization techniques that are why it is also known as the optimization based problem solving algorithm. It is also a supervised learning algorithm used for optimization of real life practices[14].

$$D = \sum ni = 1\ di\ /N$$

E. v) Support Vector Machine

SVM separates the values on the basis of hyperplane i.e. in this algorithm first we have to find the optimal hyperplane which is the best fit line that is responsible for the distribution of the dataset into multiple categories.

Here before we get the optimal hyperplane we have to find the hyperplane then the midpoint of these hyperplanes we get the optimal hyperplane and as we get the appropriate hyperplane we can easily find out the category wise data. If the values are in between the hyperplanes, then we categories according to the shortest distance with the nearest category[15].

$$L(x, a) = f(x) + \sum iaigi(x)$$

F. vi) IBK

It is java based algorithm which helps in finding accuracy. Class IBK is a machine learning algorithm from Lazy package based on weka. It stands for Instance Based K-Nearest Neighbor. It is the extended version of the k-nearest neighbor which can efficiently find out the accuracy with best results by deducting the required space which is to be used by the dataset. It is quite similar with k-nearest neighbor but it efficiently finds the value of k (a value which is unknown) which constitute the classes which are already classified between classes. It uses the same method to classify the data points as we were calculating in the k-nearest neighbor but here we are doing it in efficient way by utilizing the less space as compared with the k -nearest neighbour.[16]

distances[i] = distances[i]*distances[i]; distances[i] = Math.sqrt(distances[i]/m_NumAttributesUsed);

G. vii) J48

J48 is a classification based algorithm can be implemented on weka. J48 is a Decision Tree based algorithm which is going to generate the rules for the prediction of the target dataset. With using the tree based classification the difficult distribution of the data can easily be maintained. The objective of this algorithm is to gain a balance in terms of flexibility a nd in terms of the accuracy which matters a lot for every algorithm[17].

$$Gain(y,f) = Entropy(y - Entropy(f/y))$$

H. viii) Random Committee

In this algorithm, it finds the solution in the form of tree not in the form of data or in the form of table. It is the upgrade of the bagging for decision tree and it can be used both for regression as well as classification.

This algorithm selects the data from the dataset which is in the form of the tree it select the data in randomized f orm from the subset of the input attributes but it was not in bagging as it takes the data as much as it can handle i.e. bagging algorithm was greedy but applying this small change may affect the predictions in large amount. Results through this algorithm have the best prediction which was difficult in the bagging technique[18].

I. ix) Randomizable Filtered Classifier

This algorithm is implemented on the data which is arbitrary classified and after that the data is passed through the arbitrary filters. It works similarly like the classifier algorithm the structure of this algorithm works on the t raining data and the test instances after applying. It doesn't alter the structure of the dataset while using this algorithm each base classifier is built with a different random number seed. The final accuracy which we want comes out to be the average of the combination of all the predictions (accuracy) of each random seed [20].

J. x) Random Sub Space

Random Sub Space is a machine learning algorithm from weka. This class belongs to the Meta package. This algorithm is also known as the attribute bagging or in other ways feature bagging. It helps in reducing the correlation between the estimators by training that dataset on the random sample chosen from the dataset. It selects the f features from the dataset randomly.

In this algorithm we combine the different models produced by different learners. We do such combination because it gives the better accuracy than the other original output and the combining of these different learners is also known as the bagging[19].

$$fb(L)\ (x *\ ) = 1/\ L\ L \sum l=1\ g\hat{}\ (l)\ (x\ \square\ ).$$

K. xi) Random Forest

This algorithm is one of the most popular algorithms for finding the best accuracy for the given datasets. I t belongs to decision tree package in weka. Random forest is different from bagging only in one aspect rest all is same and that difference is it uses upgraded tree learning algorithm that selects the candidate when the tree gets split and this process sometimes is known as the feature bagging.

The reason for doing this is to correlate the tree with ordinary tree. This algorithm is a tree based algorithm which uses the tree like structure in which the dataset as per the attributes is divided into leaves/sub trees[20].

$$F = 1/B\sum Bb=1fb(x)$$

IV. RESULT

A. Graphical Representation

Graphical representation of different activities of various persons have been shown with the help of line graph. In this graph x - axis represents number of persons and y-axis represents the time taken by a person to change his body posture from one to another we have shown variation in data collected time taken by those persons to perform the activities[21].

In this graph you may wonder that why the graphs having the value in negative as time cannot be represented in negative this is because the dataset which is used in inappropriate but the results are appropriate this is due to the fact that the graph is plotted on the dataset extracted from internet but the dataset which is used for finding the accuracy was firstly pre-processed and then only was used in order to get the accuracy as per our need [22].

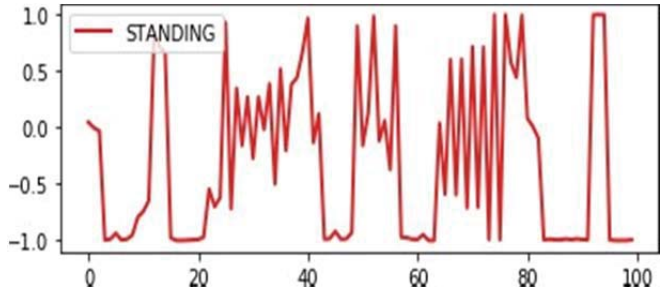

FIGURE 2. Graphical activity of standing

It shows the activity of standing i.e., how much time a person takes to change the position of his body to make stand.

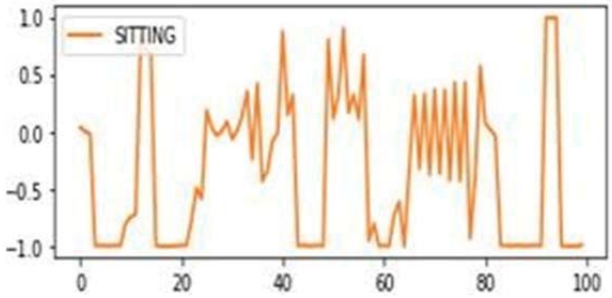

FIGURE 3. Graphical activity of sitting

It shows the activity of siting i.e., how much time a person takes to change the position of his body to make sit.

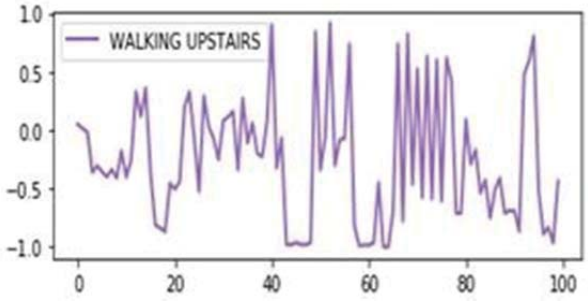

FIGURE 4. Graphical activity of walking upstairs

It shows the activity of walking upstairs i.e., how much time a person takes to change the position of his body to make walk upstairs.

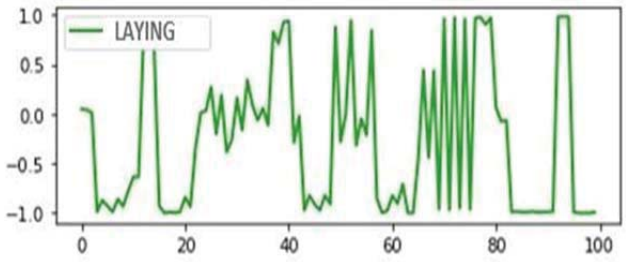

FIGURE 5. Graphical activity of laying

It shows the activity of laying i.e. how much time a person takes to change the position of his body to make lay down.

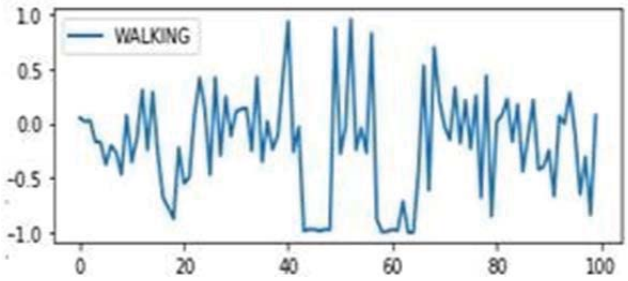

FIGURE 6. Graphical activity of walking

It shows the activity of walking i.e. how much time a person takes to change the position of his body to make start walking.

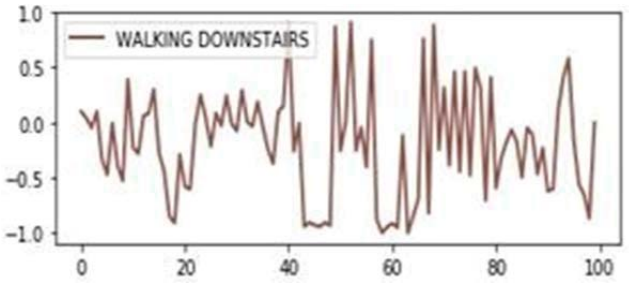

FIGURE 7. Graphical activity of walking downstairs

It shows the activity of walking downstairs i.e. how much time a person takes to change the position of his body to make walk downstairs.

Below given figure 8 is the combined graphical representation of all the human activities which are performed by humans in day-to-day life i.e. daily route these activities are involved in the daily routine of the humans which is represented separately above [23]. In the below mentioned graph x-axis represents the persons residing the different age groups and the y-axis represent the time taken by them to change the posture of the body.

The different color represents the different activities of the humans performed on the daily routine basis.

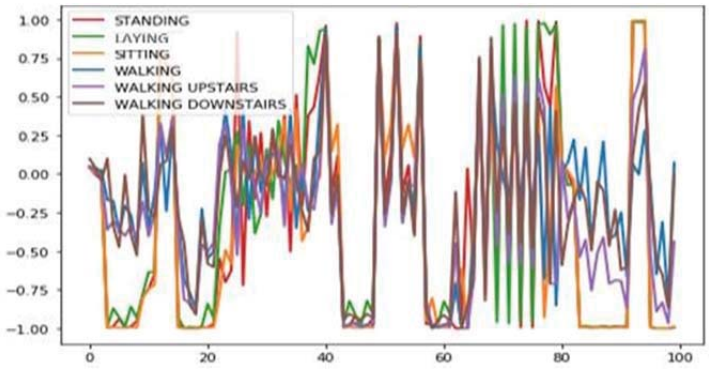

FIGURE 8. Combined gaphical activity

B. Accuracy comparison between Spyder and Weka

In this paper we are using different machine learning algorithm in order to get the best accuracy for which we have used below mentioned algorithms.

The algorithms which are used in our projects are:-

- Logistic Regression.
- Gaussian NB.
- Nearest Centroid.
- Decision Tree Classifier.
- Support Vector Machine.

In that Logistic Regression is the one which gives the highest accuracy in all of these algorithms. All of these are the best algorithm on their but in these for this particular dataset Logistic Regression is the one with best accuracy. The values which we got in the table shows how the values will get fluctuate as we change the algorithms. Here fluctuation can be termed as the change in accuracy. The Table 1 shows different algorithms and their accuracies.

In order to overcome with the problem of fluctuation which we are facing and get the accuracy which doesn't change even if we change the value of clusters then in that case we are required to use the weka software in order to

check for the different clusters used in the proposed algorithm and also to get the some new accuracies using some new algorithms in order to search for the algorithm having the better accuracy than we got in the spyder. Thus, we have to use both the software's in order to get the best accuracy as per our need and make our machine perfect for all the condition and to check the that the proposed algorithm is giving the output as per our need.

1) Accuracy Table (Spyder):

TABLE 1. Accuracy results of algorithms performed in spyder

Accuracy	Methods Used
0.945	Logistic Regression
0.747	Gaussian NB
0.815	Nearest Centroid
0.815	Decision Tree Classifier
0.918	Support Vector Machine

These algorithms are used and in that Logistic Regression is the one which gives the highest accuracy in all of these algorithms.

All of these are the best algorithm on their but in these for this particular dataset Logistic Regression is the one with best accuracy. The values which we got in the table shows how the values will get fluctuate as we change the algorithms. Here fluctuation can be termed as the change in accuracy. The below table shows about the different algorithm and their accuracies with them.

2) Accuracy Table (Weka):

TABLE 2. The above table gives the accuracies from the weka software using different algorithms.

2 cross validation	5 cross validation	10 cross validation	Methods used
0.975	0.981	0.983	IBK
0.964	0.969	0.974	J48
0.998	0.998	0.999	Random Committee
0.825	0.84	0.83	Randomizable Filtered Classifier
0.998	0.999	0.999	Random Subspace
0.958	0.959	0.961	Naïve Bayes
0.999	0.999	0.999	Random Forest

As you can see from the above table that as I changed the value of cross validation the accuracies which are prior get affected for 10 cross validation it was different, for 5 cross validation and for 2 cross validation it was different because I have decided to create the groups for the different validations and the accuracy by every algorithm got changed some of the algorithms got increased and some of them got decreased and there is only one

algorithm with which there is no effect as I change the value of cross validation and that algorithm is Random Forest and the confusion matrix for the random forest is given below:

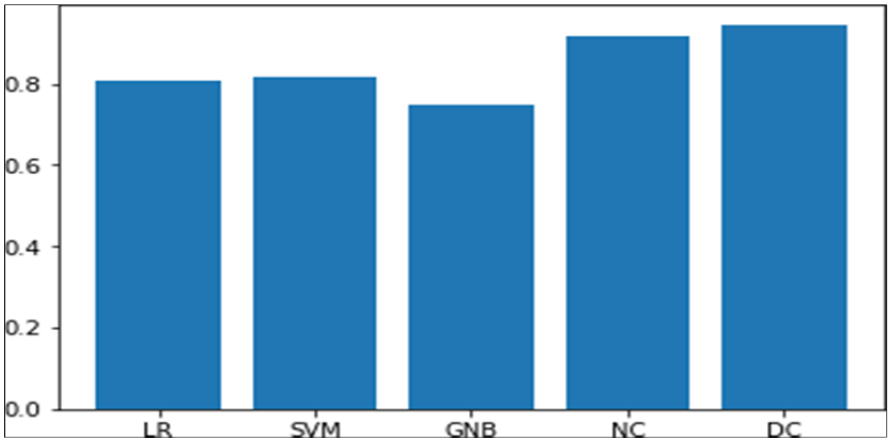

FIGURE 9. Graphical chart representation of algorithms performed in Spyder

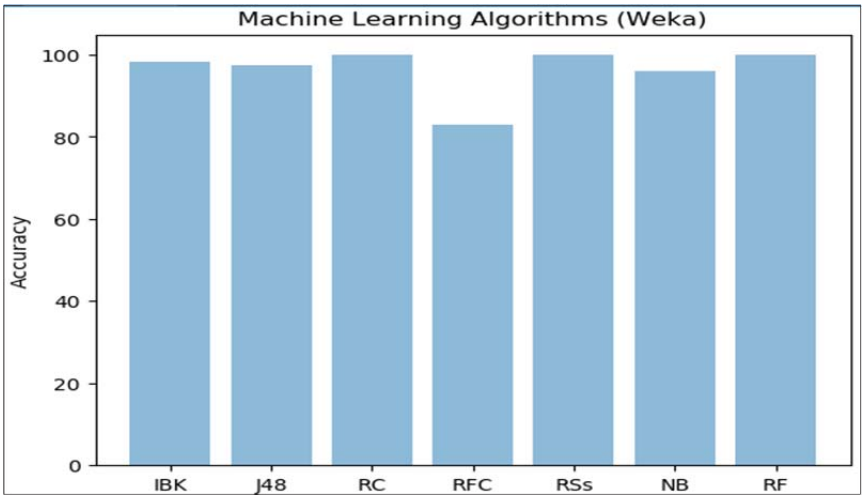

FIGURE 10. Graphical chart representation of algorithms performed in Weka

C. Cross Validation results

The below picture shows the confusion matrix for the cross validation of 10, 5 and 2 for the Random Forest algorithm which is best suited for the given dataset as the accuracy for the proposed algorithm not changed even if the groups have different number of clusters. So, with this we can say that the algorithm which we were in search is Random Forest.

Here, accuracy defines that out of 100 calculations 99.9 percent of the calculations will be observed correctly for a machine this is the accuracy which we want see.

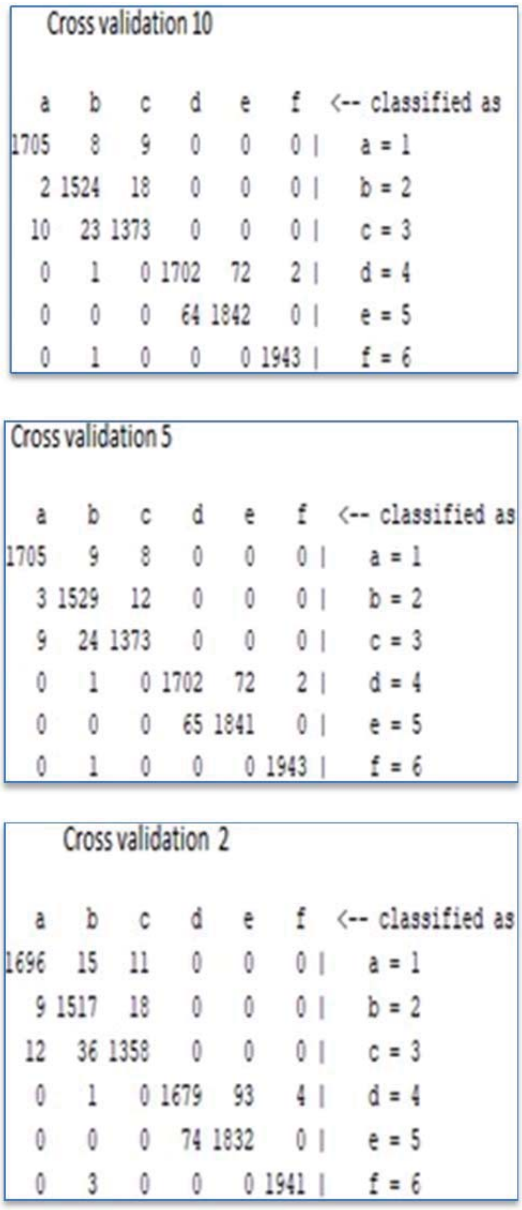

FIGURE 11. Cross validation points 5, 10 and 2

V. CONCLUSION AND FUTURE SCOPE

In this paper we examined different algorithms of machine learning in order to check which algorithm is based suited f or the human activity recognition system as in today's era it is not required to use different sensors in different body parts. As all these can be done easily through the Smartphone easily. To implement it I have used Logistic Regression, Gaussian NB, Nearest Centroid, Decision Tree Classifier Support Vector ==Machine and some other algorithms of weka and in all these I found that the algorithm which gives the maximum accuracy in Random Forest and after that it is Random Subspace. By examining this entire algorithm, we conclude that the Tree based algorithm can be best suited to implement the human activity recognition system with best accuracy.

REFERENCES

[1] C. Marinagi, P. Belsis, and C. Skourlas, "New Directions for Pervasive Computing in Logistics," *Procedia -*

Soc. Behav. Sci., vol. 73, pp. 495–502, 2013, doi: 10.1016/j.sbspro.2013.02.082.
[2] G. Chetty, M. White, and F. Akther, "Smart phone based data mining for human activity recognition," *Procedia Comput. Sci.*, vol. 46, no. Icict 2014, pp. 1181–1187, 2015, doi: 10.1016/j.procs.2015.01.031.
[3] B. Nguyen, Y. Coelho, T. Bastos, and S. Krishnan, "Trends in human activity recognition with focus on machine learning and power requirements," *Mach. Learn. with Appl.*, vol. 5, no. June, p. 100072, 2021, doi: 10.1016/j.mlwa.2021.100072.
[4] Y. Zhang, "Offline Machine Learning for Human Activity Recognition with Smartphone."
[5] Y. Min, Y. Yin, and K. Khin, "Comparing the Performance of Machine Learning Algorithms for Human Activities Recognition using WISDM Dataset," *Int. J. Comput.*, vol. 38, pp. 61–72, 2020.
[6] J. Rabbi, M. T. H. Fuad, and M. A. Awal, "Human Activity Analysis and Recognition from Smartphones using Machine Learning Techniques," 2021, [Online]. Available: http://arxiv.org/abs/2103.16490.
[7] D. Anguita, A. Ghio, L. Oneto, F. X. Llanas Parra, and J. L. Reyes Ortiz, "Human activity recognition on smartphones for mobile context awareness," *Neural Inf. Process. Syst. Conf.*, pp. 1–9, 2012, [Online]. Available: http://hdl.handle.net/2117/22167.
[8] G. De Leonardis *et al.*, "Human Activity Recognition by Wearable Sensors," *2018 IEEE Int. Symp. Med. Meas. Appl. Proc.*, p. in press, 2018, doi: 10.1109/MeMeA.2018.8438750.
[9] S. S. Dash, S. K. Nayak, and D. Mishra, "A review on machine learning algorithms," *Smart Innov. Syst. Technol.*, vol. 153, pp. 495–507, 2021, doi: 10.1007/978-981-15-6202-0_51.
[10] Z. Gulzar, A. A. Leema, and I. Malaserene, "Human Activity Analysis using Machine Learning Classification Techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 2, pp. 3252–3258, 2019, doi: 10.35940/ijitee.b7381.129219.
[11] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "Trends in human activity recognition using smartphones," *J. Reliab. Intell. Environ.*, vol. 7, no. 3, pp. 189–213, 2021.
[12] M. V.D. Prasad *et al.*, "Human activity recognition using deep learning," *Int. J. Emerg. Trends Eng. Res.*, vol. 7, no. 11, pp. 536–541, 2019, doi: 10.30534/ijeter/2019/227112019.
[13] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Energy efficient smartphone-based activity recognition using fixed-point arithmetic," *J. Univers. Comput. Sci.*, vol. 19, no. 9, pp. 1295–1314, 2013.
[14] L. B. Marinho, A. H. de Souza Junior, and P. P. R. Filho, "A new approach to human activity recognition using machine learning techniques," *Adv. Intell. Syst. Comput.*, vol. 557, no. February, pp. 529–538, 2017, doi: 10.1007/978-3-319-53480-0_52.
[15] S. O. Slim, A. Atia, M. M. A. Elfattah, and M. S. M. Mostafa, "Survey on human activity recognition based on acceleration data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 3, pp. 84–98, 2019, doi: 10.14569/IJACSA.2019.0100311.
[16] H. Vellampalli, "Physical Human Activity Recognition Using Machine Learning Algorithms A dissertation submitted in partial fulfillment of the requirements of," 2017.
[17] F. R. Allen, E. Ambikairajah, N. H. Lovell, and B. G. Celler, "Classification of a known sequence of motions and postures from accelerometry data using adapted Gaussian mixture models," *Physiol. Meas.*, vol. 27, no. 10, pp. 935–951, 2006, doi: 10.1088/0967-3334/27/10/001.
[18] S. Grijalva, G. Cueva, D. Ramírez, and W. G. Aguilar, "HSVM-Based Human Activity Recognition Using Smartphones," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11744 LNAI, pp. 217–228, 2019, doi: 10.1007/978-3-030-27541-9_19.
[19] M. Z. Uddin and A. Soylu, "Human activity recognition using wearable sensors, discriminant analysis, and long short-term memory-based neural structured learning," *Sci. Rep.*, vol. 11, no. 1, pp. 1–15, 2021, doi: 10.1038/s41598-021-95947-y.
[20] S. Oh, A. Ashiquzzaman, D. Lee, Y. Kim, and J. Kim, "Study on human activity recognition using semi-supervised active transfer learning," *Sensors*, vol. 21, no. 8, 2021, doi: 10.3390/s21082760.
[21] Ankita, S. Rani, H. Babbar, S. Coleman, A. Singh, and H. M. Aljahdali, "An efficient and lightweight deep learning model for human activity recognition using smartphones," *Sensors*, vol. 21, no. 11, pp. 1–17, 2021, doi: 10.3390/s21113845.
[22] V. Mittal, D. Umrania, P. Bansal, and P. Ratnani, "Modelling and Analysis of Various Machine Learning Algorithms for Human Activity Recognition," *SSRN Electron. J.*, no. Icicnis, 2021, doi: 10.2139/ssrn.3883148.
[23] S. Ramasamy Ramamurthy and N. Roy, "Recent trends in machine learning for human activity recognition—A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, 2018, doi: 10.1002/widm.1254.