International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Prediction of Human Physiological States by Using an Enhanced Recursive Feature Elimination Method

Thangapriya[1], Nancy Jasmine Goldena[2]

[1]Research Scholar, Registration No:20211242282010, Department of Computer Applications and Research Centre, Sarah Tucker College (Autonomous), Affiliated to Manonmaniam Sundaranar University, Abhishekapatti, Tirunelveli-627012

[2]Associate Professor, Department of Computer Applications and Research Centre, Sarah Tucker College (Autonomous), Affiliated to Manonmaniam Sundaranar University, Abhishekapatti, Tirunelveli-627012

Abstract

Human Action Recognition (HAR) is a vital part of the healthcare sector. Medical practitioners are experiencing difficulty in recognizing human physiological conditions due to the enormous quantity of sensory stimulation. Machine learning techniques help to predict human physiological conditions, providing medical practitioners to work more efficiently. Feature selection is an essential part of discovering new knowledge in the majority of real-world problems when there are a lot of features. Feature selection is extremely useful because it speeds up decisions and enhances classification performance. The importance of feature selection in machine learning is dimension reduction in a massive multivariate data collection. This paper presents an effective feature selection method known as the Enhanced Recursive Feature Elimination (EFRE) for selecting key features from the data set for HAR prediction. The experimental results reveal that the ERFE technique selects the most suitable features for HAR prediction. The performance of the proposed ERFE approach is tested using different performance evaluation metrics. The performance analysis shows that the ERFE method outperforms existing feature selection methods with 88% accuracy.

Keywords: Activity recognition, Classification, ERFE, Feature Selection, Performance Evaluation

1. Introduction:

Health is essential for living a fulfilling life. Healthy living keeps the body and the mind active. Research in the medical field analyses recent information that leads to the development of new treatments, vaccinations, healthcare devices, and procedures, as well as assist in improving existing treatment procedures. HAR is an intensive research area for decades. HAR is a main research field in the medical science firm that provides information on a person's physiological status [4]. Machine Learning (ML) is always used to solve real-time problems. The majority of existing ML approaches are used for predicting basic human behaviours such as walking, running, sitting, sleeping, eating, and so on. However, this work tries to predict human physiological conditions such as mental, emotional, physical, and neutral.

The process of selecting a subset from the set of features with a clearly specified performance measure is

known as Feature Selection (FS) or feature reduction [1]. FS is a commonly used method for dimensionality reduction that eliminates unnecessary and duplicate features [2]. FS increases classifier's efficiency, and simplifies the model [3]. Health monitoring devices such as the electrocardiogram (ECG), thoracic electrical bioimpedance (TEB), and electro dermal activity (EDA meter) are used to collect data [8]. These devices are used to obtain physiological information from people and the data obtained is in the form of a signal [7]. The signal is digitized and its features are extracted [5]. The analysis is then carried out using these derived features.

The activity dataset is taken from the UCI machine learning repository [25]. A total of 532 features have been extracted from the raw signal, including 4480 values available for modelling. The activity dataset, consists of 4480 rows and 533 columns, the 533th column indicates the physiological condition of a person which is emotional, mental, physical, or neutral.

2. Related Works:

HAR research is becoming progressively significant in HCI, healthcare monitoring, and elderly support. Based on the sensors used, HAR systems are classified into three types. They are multi-sensor fusion, device-free, and device-based. Multi-sensor fusion of data methodology has been widely used in object recognition, smart home appliances, automation, medical care, image analysis, pattern recognition, and various other applications. The use of sensors is growing rapidly due to technological advancements. The camera can supply visual information for device-free activity recognition. Cameras are used by researchers to recognize behavioural patterns of individuals, walking directions, identify suspected targets, detect possible suspicious actions in advance, and secure resources that are at risk of intrusions. Device-based HAR methods need individuals to wear sensors on particular areas of their bodies so the sensors can record information about their actions.

A HAR model was presented by L. Fang et al. to identify human actions, particularly the actions of up-and-down buses. The effectiveness of using a smartphone sensor to identify activities using a k-NN classifier with RFE feature selection method has been demonstrated using the Matlab and Weka software platforms. This useful existing work explains how to use public resources to deal with personal everyday safety while travelling and also saving energy [22]. To identify human regular behaviours, Y.-L. Hsu et al. proposed a wearable human activity categorization system based on inertial sensing. The proposed system collected signals from motion of human activities using two inertial recognising parts worn on the participants' wrist and the ankle. Using the Non-parametric Weighted Feature Extraction (NWFE) approach, the user can effectively decrease the number of feature size and increase classification rate [23]. A feature selection method called BPSO was proposed by Y. Chen et al. to enhance the efficiency of healthcare data categorization. The proposed strategy has been tested against a number of benchmark systems and validated using a number of UCI public datasets [24].

3. Existing Feature Selection Techniques:

When classifying the activities, not all features are important equally in machine learning and every aspect must be considered. To avoid overfitting caused by unnecessary dimensions, an intelligent FS technique is needed to discover the significant features [6]. The FS algorithm must focus on the important subset of features while ignoring the rest. To choose the optimal performance feature subset for incorporation into the testing set, the specific FS algorithm uses the training set as a test platform. The ERFE FS method is used to reduce the dimensionality of the original feature space in order to achie-

ve a better classification result.

3.1 LASSO Method:

LASSO is an acronym that stands for Least Absolute Shrinkage and Selection Operator. LASSO is a supervised regularization approach used in ML that utilizes shrinkage, where the data values are shrunken towards a central point such as the mean value [17]. LASSO FS method is used to remove redundant features to select best features [15]. LASSO reduces coefficients to absolute zero; LASSO's L1 regularization reduces the regression coefficient for the lowest contributing variable to zero or close to zero [19]. For the LASSO approach, the following equation is used, where x = predictor variable, y = response variable, ʎ = tuning parameter

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \alpha\sum_{J=1}^{m}|\beta_j| \qquad (1)$$

3.2 Random Forest for Valuable Selection Method:

Random Forest (RF) is a wrapper FS method that works by training a large number of decision trees and classifying the activity classes. RF is a common technique for simplifying data preparation and providing faster responses [20]. Out of all the predictor variables, some are chosen at random, as well as the result obtained on these parameters is utilized to split the node [21]. Based on Gini impurity or information gain techniques, the optimal split is selected. Algorithms with built-in FS methods are used to implement them.

3.3 Recursive Feature Elimination Method:

A popular approach for choosing features that are most important for predicting the input variables in a predictive model for classification is called Recursive Feature Elimination (RFE) [13]. To identify the best possible combination of attributes, RFE uses a backward selection procedure [14]. The process begins by creating a model based on all features determining the relative relevance of each feature in the model [15]. By using proper evaluation metrics RFE removes the features which has the least importance [16]. By this way the feature importance is calculated. This procedure is repeated until only a smaller subset of the features is remained. The resultant model is then trained using the best subset.

4. Outline of the Work:

The dataset that includes human physiological features is pre-processed to identify missing values and remove unnecessary information. The pre-processed data will be the input into the proposed ERFE FS algorithm. The attributes selected are fed into the classifier so that it can learn. The prediction process in this work is based on supervised learning. The classifier is trained on training samples, then samples that are unidentified are given to verify the newly constructed classifier. In order to select the suitable activity class, the outcomes are then assessed using certain performance indicators. The outline of the work is shown in Figure 1.

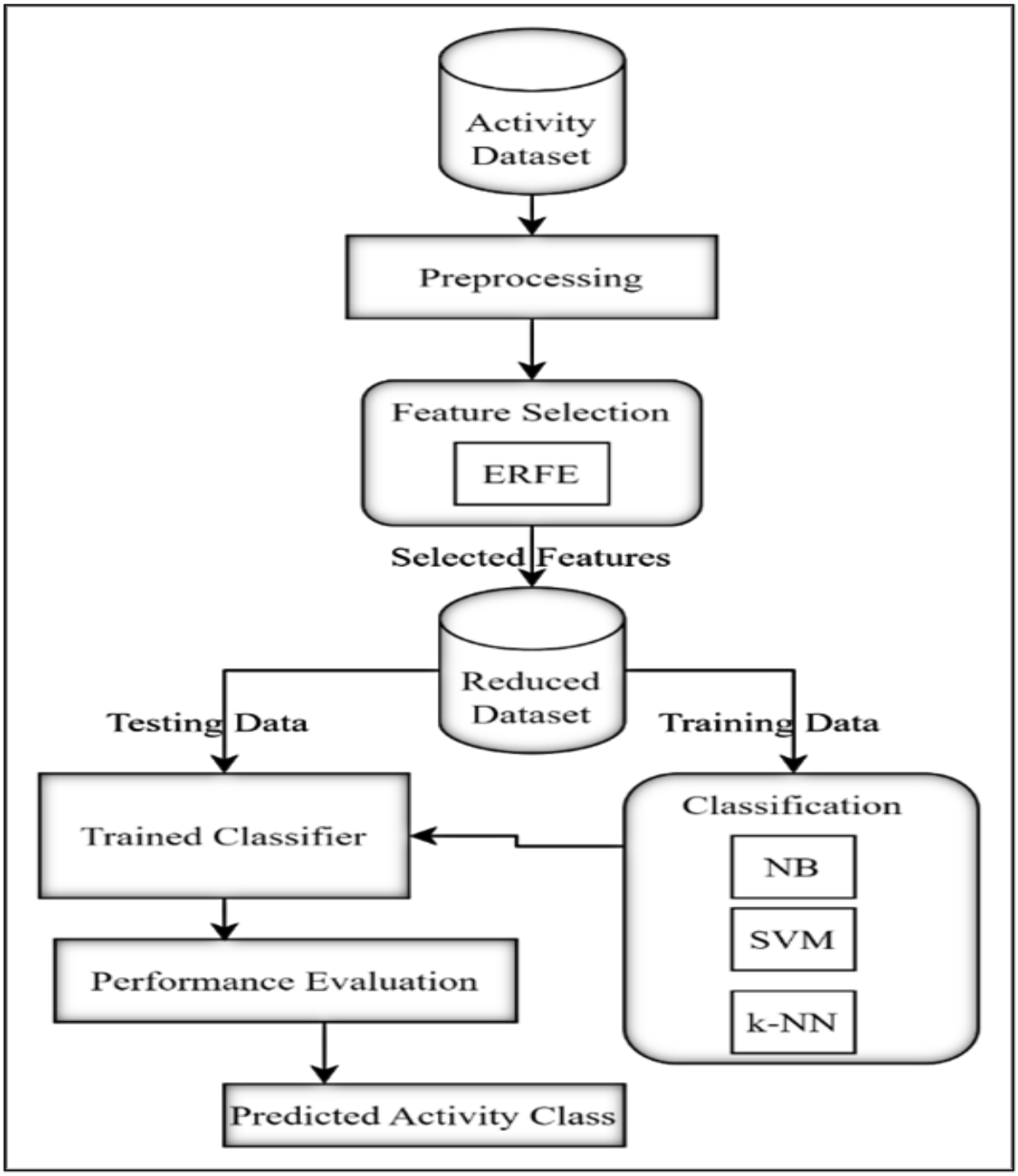


Figure 1: Outline of the Work

5. Proposed Methodology:

5.1. Enhanced Recursive Feature Elimination (ERFE) Method:

The existing RFE technique eliminates the least significant features by ranking method. The remaining features has an effect on classification accuracy [12]. For feature selection, existing RFE method requires a permutation procedure under brute force approach. Removing weak features takes excessive computational time because updating the reduced dataset is the most challenging aspect of the RFE [11]. The ERFE approach is proposed to overcome the drawbacks of the existing RFE feature selection method [10]. It demonstrates that the work improves classification results so that the best features from

the activity dataset can be selected through the ERFE. ERFE enables the machine learning system to train more quickly. The outline of the proposed research is shown in Figure 2.

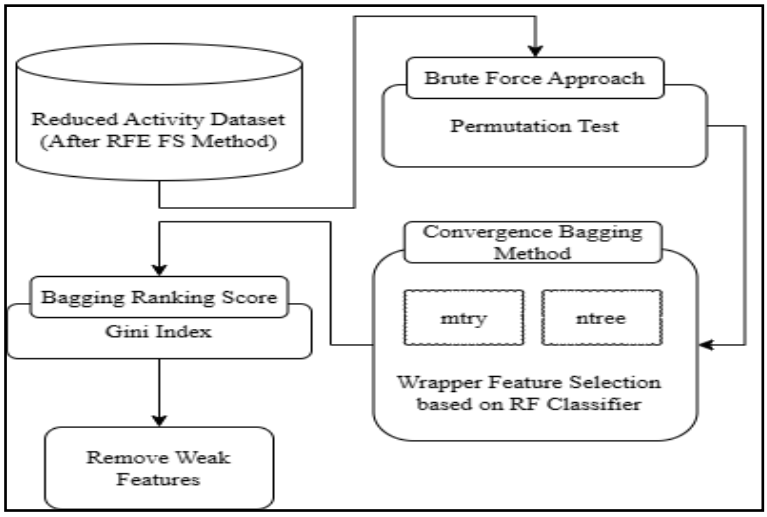


Figure 2: Architecture of Proposed ERFE

5.2 The following steps are used to implement the proposed ERFE FS Method:

Step 1: Load the activity dataset.

Step 2: For the purpose of locating the most useful aspects, RFE employs a brute force approach. The activity dataset is represented as a matrix with dimensions of n x m, where n represents rows and m represents columns. Following the reorganization, the input activity dataset is merged with the modified dataset.

The original dataset and the combined activity dataset are then mixed together, and the shuffling procedure is used to create the expanded dataset. This dataset is also referred as permuted activity dataset. This dataset can be utilized to determine the significance of the features. The standard deviation value decreases as the dataset are expanded, which indicates that the value is getting closer to the average.

Step 3: The minimum value, first quartile, median, mean, third quartile, and maximum value are all basic statistical values computed by the summary() function for further evaluation.

Step 4: RF classifier is used to find the most essential features from the permuted dataset. The RF classifier generates a set of decision trees from a randomly chosen section from the permuted activity data set. The permuted activity dataset is fed into the RF classifier (basically a collection of decision trees). When RF operates on numerous decision trees, it can estimate missing values more accurately, runs on an expanded data set, and has a lower probability of over-fitting. Because of these special qualities, the proposed ERFE feature selection outperforms the existing RFE feature selection.

The overall process of this step is referred as convergence bagging method. The convergence bagging method is applied in order to reduce the number of distinct features. The final performance of the generalization is calculated by taking the average of the performance acquired individually for each reduction level. The following are the RF classifier's parameters:

1. The number of decision trees or ntree: It refers to the number of trees that will be established and the standard tree size is 500. Increasing or decreasing the number of trees has no effect on the outcomes.
2. The number of variable splits or mtry: The number of variables that will be included in the first split when developing a tree is called mtry. The root square of the features is the ideal value for the mtry.

3. nodesize or minimum node size: The minimal number of observations that should be included in a terminal node is indicated by this parameter.

Step 5: The absolute bagging score is the standard score used to compare the importance of the features chosen using the convergence bagging method. The bagging ranking score is calculated by taking the average of the weights for each of the different levels in the second split. The final set of the reduction level is obtained by merging the values with the highest absolute score. The following formula is used to determine the absolute bagging score, where X represents the single value in the raw data, μ represents the average of all values, σ represents the standard deviation.

$$\textbf{Absolute Bagging Score} = z = \frac{(X-\mu)}{\sigma} \qquad (2)$$

Step 6: To eliminate the weak features from the dataset, the Gini index is computed. The probability that a randomly selected instance will be incorrectly classified is measured by the Gini Index, also known as Impurity. The chance of misclassification is better when the Gini Index is lower. This Gini index re-rank approach is used to rank each feature from most effective to least. Adding this step strengthens the FS approach. The formula below is used to get the Gini index, where P(i) is the total number of observations in the node, and j is the number of classes in the target variable.

$$Gini = 1 - \sum_{i=1}^{j} P(i)^2 \qquad (3)$$

Step 7: Based on the optimization method find the best optimal features.

6. Classification Methods:

After the selection of features, the three supervised machine learning techniques such as Support Vector Machine (SVM), Naive Bayes (NB), and k-Nearest Neighbor (k-NN) are utilized to further classify the physiological conditions.

6.1 HAR Classification Procedure:

The basic procedure for solving the HAR prediction problem using the proposed ERFE FS approach with existing classifier is presented below.

Step 1: The input dataset is an activity dataset containing extracted features from activity tracking devices.

Step 2: The input dataset is preprocessed for abnormalities such as missing values and redundant data.

Step 3: The proposed ERFE FS approach is used to pick significant features from the preprocessed data.

Step 4: The features that have been chosen are used in the prediction procedure. The reduced dataset is divided into two parts for training and testing purposes.

Training Phase

Step 5: A total of 70% of the samples from the given dataset are selected to operate as training samples.

Step 6: During the learning phase, the classification algorithm is applied to all of the training data.

Step 7: Using the training dataset for HAR prediction, the algorithm is well-trained.

Testing Phase

Step 8: Testing samples are selected from the remaining 30% of the given dataset.

Step 9: The testing samples are used to apply the existing classifier to the prediction process.

Step 10: The trained existing classification method predicts an activity class and identifies the target class for the newly provided input.

7. Performance Evaluation Result

The performance test is conducted to evaluate the proposed ERFE with various existing classification methods. ERFE enables the ML classifiers to train more quickly than other FS methods. ERFE reduces the model's complexity and simplifies interpretation.

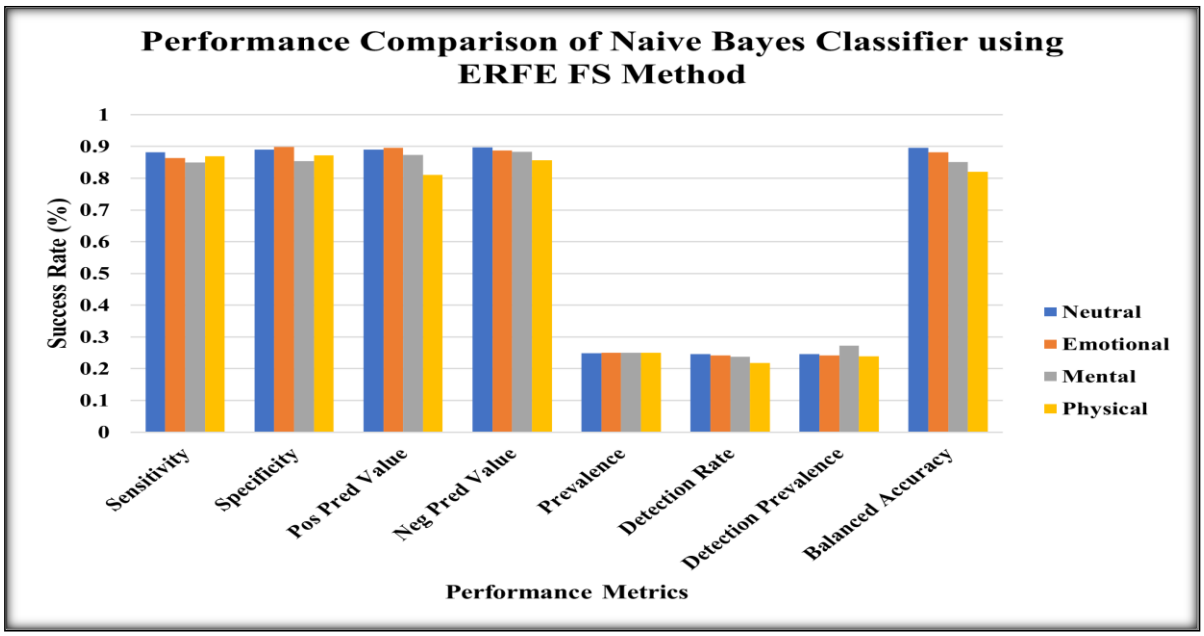


Figure 3: Performance Evaluation Result for Naive Bayes Classifier using ERFE Feature Selection Method

To improve the efficiency of the MRFE approach, the RF classifier's parameters are fine-tuned to find primary features. The selected features' (feature subset) importance is calculated by varying the range of the RF parameters, ntree and mtry, while analysing the performance.

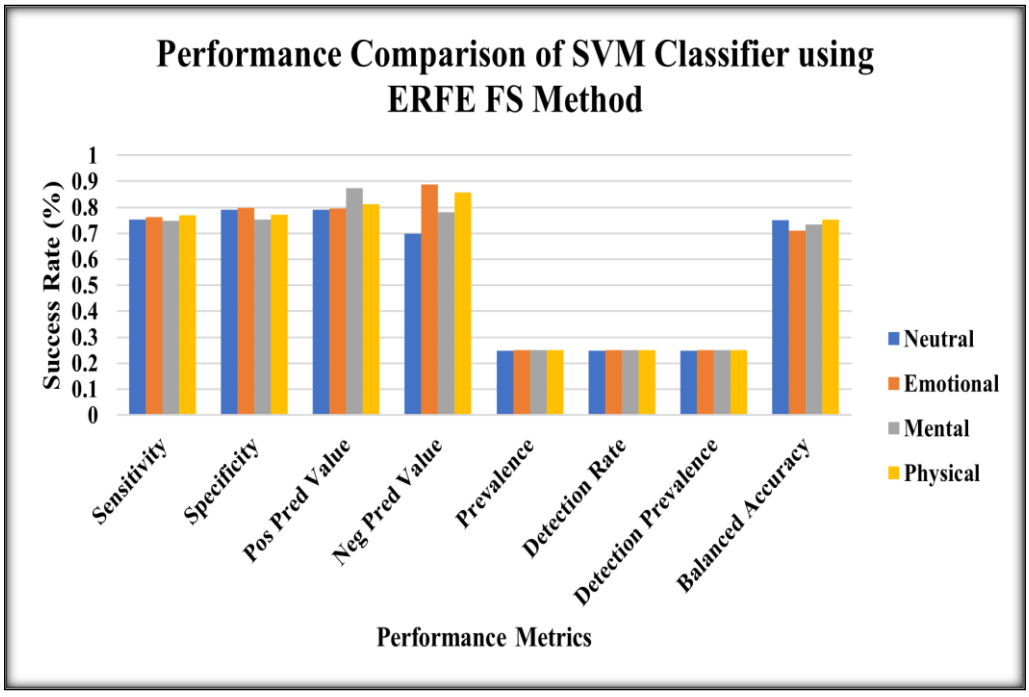


Figure 4: Performance Evaluation Result for Support Vector Machine Classifier using ERFE Feature Selection Method

Additionally, ERFE prevents overfitting and maximizes the accuracy of the model while selecting the most suitable subset. The enhanced method provides the optimum response for over fitting and redundant features, and it is perfect for predictive modelling, particularly for FS in the HAR problem.

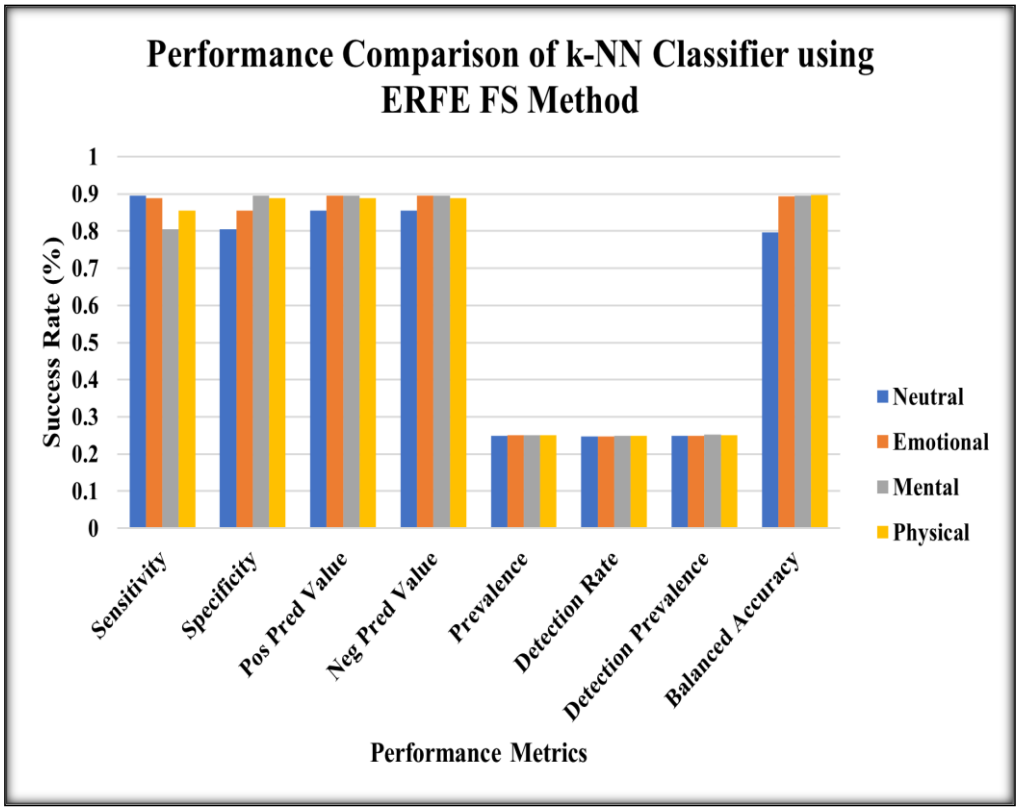
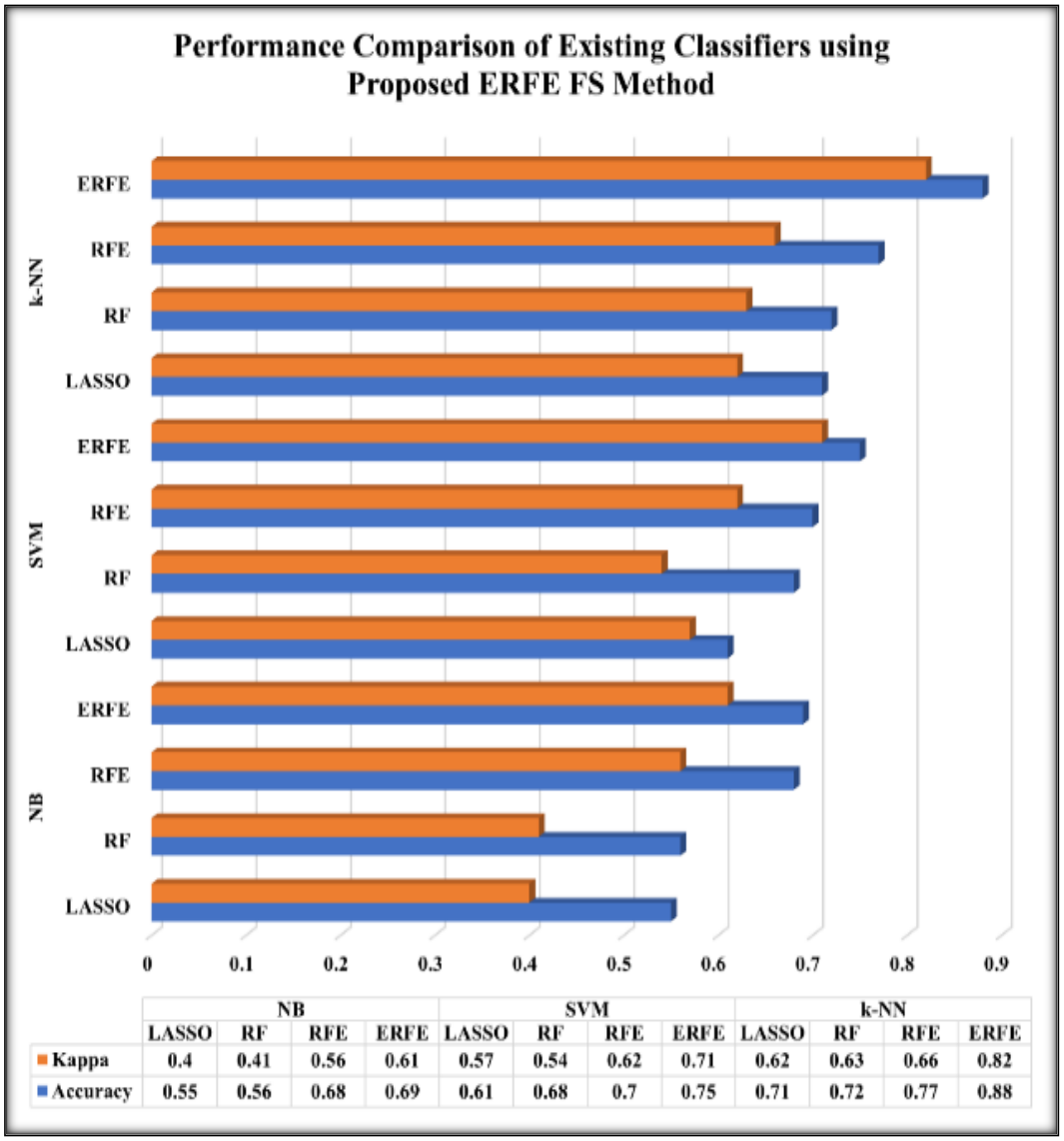


Figure 5: Performance Evaluation Result for k-Nearest Neighbour Classifier using ERFE Feature Selection Method

Out of the 532 input features, 239 features are selected by the LASSO FS method, 378 features are selected by the RF FS method, 155 features are selected by RFE, and 105 features are selected by ERFE. After the selection of features, the three supervised ML methods such as NB, SVM and k-NN are utilized to further classify whether a person belongs to emotional, mental, physical, or neutral states. The optimum outcome is revealed by applying performance evaluation metrics like accuracy and kappa.

The performance result of the proposed ERFE FS method with existing classification methods are evaluated. Sensitivity, specificity, positive prediction value, negative prediction value, prevalence, detection rate, detection prevalence and balanced accuracy are calculated for various existing classifiers using proposed ERFE method. Figure 3 shows the performance evaluation result for the Naive Bayes classifier using the ERFE feature selection method for each class. Figure 4 shows the performance evaluation results for the support vector machine classifier using the ERFE feature selection method for each class.



	NB				SVM				k-NN			
	LASSO	RF	RFE	ERFE	LASSO	RF	RFE	ERFE	LASSO	RF	RFE	ERFE
Kappa	0.4	0.41	0.56	0.61	0.57	0.54	0.62	0.71	0.62	0.63	0.66	0.82
Accuracy	0.55	0.56	0.68	0.69	0.61	0.68	0.7	0.75	0.71	0.72	0.77	0.88

Figure 6: Overall Performance Evaluation Result of Existing Classifiers using Proposed ERFE Feature Selection Method

Figure 5 shows the performance evaluation result for the k-nearest neighbour classifier using the ERFE feature selection method for each class. Figure 6 depicts the overall performance evaluation result of existing classifiers with the proposed ERFE feature selection method for each class. Finally, when compared to other approaches, ERFE with k-NN classification provides the best results.

8. Conclusion and Future Scope:

HAR is an important area of study in medical research. Predicting human physiological states is extremely helpful in the healthcare sector since it can indicate a particular person's health status. In this paper, the ERFE, a unique approach, is proposed for selecting the key features from the activity dataset utilizing a brute-force approach and a convergence bagging method. Experiments were done to assess the effectiveness of the proposed ERFE method for HAR prediction using k-NN, NB, and SVM classification methods. The ERFE technique is evaluated and compared to other methods, including LASSO, RF, and RFE, using activity dataset. The ERFE with k-NN classifier produces the best results among the existing approaches according to performance evaluation results.

In this work, human's emotional, mental, physical, and neutral activity classes are predicted. In future, more distinct activities can be predicted, and data from various human tracking devices can be collected. The activity may take on various forms; it may involve sexual activity or any other kind of action which can be used to protect women's safety.

Acknowledgment
We are thankful to Tamil Nadu State Council for Science and Technology for the Financial Support under RFRS (Research Fund for Research Scholar) Scheme.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
1. Fan, Changjun, and Fei Gao. "Enhanced Human Activity Recognition Using Wearable Sensors via a Hybrid Feature Selection Method." Sensors, vol. 21, no. 19, MDPI AG, Sept. 2021, p. 6434. Crossref, https://doi.org/10.3390/s21196434.
2. Li, Frédéric, et al. "Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors." Sensors, vol. 18, no. 3, MDPI AG, Feb. 2018, p. 679. Crossref, https://doi.org/10.3390/s18020679.
3. Cilia, Nicole Dalia, et al. "Comparing Filter and Wrapper Approaches for Feature Selection in Handwritten Character Recognition." Pattern Recognition Letters, vol. 168, Elsevier BV, Apr. 2023, pp. 39–46. Crossref, https://doi.org/10.1016/j.patrec.2023.02.028.
4. A. Badawi, A. Al-Kabbany and H. Shaban, "Daily Activity Recognition using Wearable Sensors via Machine Learning and Feature Selection," 2018 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2018, pp. 75-79, doi: 10.1109/ICCES.2018.8639309.
5. Wang, Aiguo, et al. "Evaluation of Random Forest for Complex Human Activity Recognition Using Wearable Sensors." 2020 International Conference on Networking and Network Applications (NaNA), IEEE, Dec. 2020. Crossref, https://doi.org/10.1109/nana51271.2020.00060.
6. Badawi, Abeer A., et al. "Sensor Type, Axis, and Position-Based Fusion and Feature Selection for Multimodal Human Daily Activity Recognition in Wearable Body Sensor Networks." Journal of Healthcare Engineering, vol. 2020, Hindawi Limited, June 2020, pp. 1–14. Crossref, https://doi.org/10.1155/2020/7914649.
7. Ahmed, Nadeem, et al. "Enhanced Human Activity Recognition Based on Smartphone Sensor Data Using Hybrid Feature Selection Model." Sensors, vol. 20, no. 1, MDPI AG, Jan. 2020, p. 317. Crossref, https://doi.org/10.3390/s20010317.
8. Mohino-Herranz, Inma, et al. "Activity Recognition Using Wearable Physiological Measurements: Selection of Features From a Comprehensive Literature Study." Sensors, vol. 19, no. 24, MDPI AG, Dec. 2019, p. 5524. Crossref, https://doi.org/10.3390/s19245524.
9. Badshah, Mustafa. Sensor - Based Human Activity Recognition Using Smartphones. San Jose State University Library. Crossref, https://doi.org/10.31979/etd.8fjc-drpn.
10. Othman, N.A., Aydin, I. (2021). Challenges and limitations in human action recognition on unmanned aerial vehicles: A comprehensive survey. Traitement du Signal, Vol. 38, No. 5, pp. 1403-

1411. https://doi.org/10.18280/ts.380515
11. Alzahrani, Mona Saleh and Salma Kammoun. “Human Activity Recognition: Challenges and Process Stages.” International Journal of Innovative Research in Computer and Communication Engineering 2016 (2016): n. pag.
12. Sunny, Jubil T et al. “Applications and Challenges of Human Activity Recognition using Sensors in a Smart Environment.” (2015).
13. Available from: https://www.linkedin.com/pulse/what-recursive-feature-elimination-amit-mittal
14. Available from: https://bookdown.org/max/FES/recursive-feature-elimination.html
15. Available from: https://machinelearningmastery.com/rfe-feature-selection-in-python
16. Available from: https://topepo.github.io/caret/recursive-feature-elimination.html
17. Available from: https://towardsdatascience.com/feature-selection-in-machine-learning-using-lasso-regression
18. Available from: https://medium.com/@23.sargam/lasso-regression-for-feature-selection-8ac2287e25fa
19. Available from: https://corporatefinanceinstitute.com/resources/knowledge/other/lasso
20. Available from: https://chrisalbon.com/code/machine_learning/trees_and_forests/feature_selection_using_random_forest
21. Available from: https://blog.datadive.net/selecting-good-features-part-iii-random-forests
22. L. Fang, S. Yishui and C. Wei, "Up and down buses activity recognition using smartphone accelerometer," 2016 IEEE Information Technology, Networking, Electronic and Automation Contrl Conference, Chongqing, China, 2016, pp. 761-765, doi: 10.1109/ITNEC.2016.7560464.
23. Y. -L. Hsu, S. -L. Lin, P. -H. Chou, H. -C. Lai, H. -C. Chang and S. -C. Yang, "Application of nonparametric weighted feature extraction for an inertial-signal-based human activity recognition system," 2017 International Conference on Applied System Innovation (ICASI), Sapporo, Japan, 2017, pp. 1718-1720, doi: 10.1109/ICASI.2017.7988270.
24. Y. Chen, Y. Wang, L. Cao and Q. Jin, "CCFS: A Confidence-Based Cost-Effective Feature Selection Scheme for Healthcare Data Classification," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 3, pp. 902-911, 1 May-June 2021, doi: 10.1109/TCBB.2019.2903804.
25. Activity recognition using wearable physiological measurements. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5RK6V