

Improving time efficiency in big data through progressive sampling-based classification model

Nandita Bangera^{1,2}, Kayarvizhy², Shubham Luharuka³, Asha S Manek⁴

¹Department of Computer Science and Engineering, RV Institute of Technology and Management, Bangalore, India

²Department of Computer Science and Engineering, B.M.S College of Engineering, Visvesvaraya Technological University, Belagavi, India

³Xvigil, CloudSEK Information Security Pvt. Ltd, Bangalore, India

⁴Department of Computer Science and Engineering, T. John Institute of Technology, Bangalore, India

Article Info

Article history:

Received Jul 3, 2023

Revised Sep 29, 2023

Accepted Oct 25, 2023

Keywords:

Classification

Feature importance

Progressive sampling

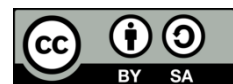
Random forest regressor

Time complexity

ABSTRACT

The proposed system aims to overcome challenges posed by large databases, data imbalance, heterogeneity, and multidimensionality through progressive sampling as a novel classification model. It leverages sampling techniques to enhance processing performance and overcome memory restrictions. The random forest regressor feature importance technique with the gini significance method is employed to identify important characteristics, reducing the data's features for classification. The system utilizes diverse classifiers such as random forest, ensemble learning, support vector machine (SVM), k-nearest neighbors' algorithm (KNN), and logistic regression, allowing flexibility in handling different data types and achieving high accuracy in classification tasks. By iteratively applying progressive sampling to the dataset with the best features, the proposed technique aims to significantly improve performance compared to using the entire dataset. This approach focuses computational resources on the most informative subsets of data, reducing time complexity. Results show that the system can achieve over 85% accuracy even with only 5-10% of the original data size, providing accurate predictions while reducing data processing requirements. In conclusion, the proposed system combines progressive sampling, feature selection using random forest regressor feature importance (RFRFI-PS), and a range of classifiers to address challenges in large databases and improve classification accuracy. It demonstrates promising results in accuracy and time complexity reduction.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Asha S Manek

Department of Computer Science and Engineering, T. John Institute of Technology

Bangalore, India

Email: ashasm.cse@gmail.com

1. INTRODUCTION

The exponential growth of data in various domains has led to the emergence of big data analytics [1] as a critical research area. Machine learning algorithms, particularly in the domains of classification, regression, and clustering, play a vital role in analyzing and understanding these data models. However, working with big data poses significant challenges. The size and complexity of datasets often lead to high computational demands, making traditional analytical methods inefficient and time-consuming. Additionally, the inclusion of irrelevant or redundant variables in the data can hinder the performance and accuracy of machine learning models. To overcome these challenges, efficient pre-processing techniques and feature selection methods have become essential in the field of big data analytics. One of the efficient methods to

reduce the dataset volume is sampling. Different sampling techniques have been implemented across various big data analytics. Though Sampling may result in handling data efficiently but it may lack the relevant information needed for classification due to its heterogeneous, imbalance nature. So it is necessary to choose the efficient preprocessing methods before the sampling process and the right kind of the sampling to reduce the issue of computational time.

One prominent approach to address the time complexity issue and improve the efficiency of classification models is progressive sampling (PS). Progressive sampling [2] is a technique that involves selecting samples iteratively for classification and prediction on massive datasets. By gradually sampling the data, this approach aims to reduce the computational burden while still maintaining high accuracy. Progressive sampling works by selecting a portion of the data for training the model iteratively, gradually expanding the size of the training set as the model improves.

In sampling, progressive techniques optimize the selection of representative data points for analysis or training. One such method is progressive sampling, which aims to achieve efficient sampling while reducing computation time. PS is likely employing iterative strategies to adaptively select new samples based on information gained from previously selected samples. The relationship between sample number and model accuracy is often represented by a learning curve, as shown in Figure 1. Learning curves offer insights into how the amount of training data impacts model performance. However, it's important to note that learning curves can exhibit different behaviors depending on the context and dataset. Once a learning curve reaches its ultimate plateau, it indicates model convergence, where optimal performance is achieved, and further iterations or additional training samples may not significantly improve accuracy. Identifying the convergence point is crucial for determining when to stop training or assess model performance.

Progressive techniques, including progressive sampling, provide efficient ways to enhance accuracy and efficiency in machine learning algorithms. Producing a representative sample from data collection involves fulfilling two important criteria. Firstly, the sample should exhibit informational equivalence by capturing and preserving the same amount of information as the complete dataset. This ensures that the learning system can extract similar insights and make accurate predictions or inferences based on the sample alone. Achieving informational equivalence is crucial for generalization and maintaining the model's performance on unseen data. Secondly, the sample should be minimized in size while still maintaining its representativeness. Reducing the sample size offers advantages such as decreased computational requirements, storage space, and processing time. A smaller sample facilitates more efficient learning and analysis, particularly with large datasets. However, it's essential to strike a balance between reducing sample size and preserving critical information. The sample should be sufficient to capture the essential characteristics and patterns of the complete dataset without compromising representativeness.

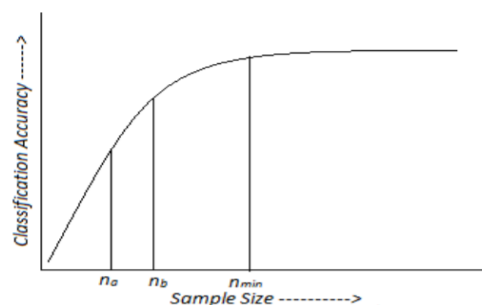


Figure 1. Progressive learning curve

Here's a high-level overview of how progressive sampling is used to reduce time complexity in a classification model for big data:

Step 1: split the data into smaller subsets.

Step 2: randomly select a lesser portion of the data, often referred to as the "seed" dataset.

Step 3: train a classification model using the seed dataset.

Step 4: evaluate the performance of the trained model using a validation dataset or cross-validation techniques.

Step 5: gradually increase the size of the training set by incorporating additional subsets of the data.

Step 6: update the existing model using the new training data.

Step 7: re-evaluate the model's performance.

Step 8: repeat steps 5-7 until the termination condition is met.

– Related work

Progressive sampling iteratively improves model accuracy and reduces time complexity by incorporating informative data subsets. It optimizes computational resources by avoiding processing the entire big data set. Considerations like data stratification [3] examining the effect of quality improvement initiatives on decreasing racial disparities in maternal morbidity, class distribution balance [4], and concept drift handling are important in designing and implementing progressive sampling. Many research techniques have used progressive sampling in different dimensions. Umarani and Punithavalli [5], explored progressive sampling approach for frequent datasets for association rule mining. Though this technique yielded improved results, the selection of dataset was limited to a particular category. Chen and Zeng [6], tackles the joint problem of automatic model selection and feature selection in machine learning. The authors propose a progressive sampling-based approach that selects both the optimal machine learning model and features simultaneously. Their framework iteratively chooses samples for model training and feature selection, gradually increasing the sample size while evaluating different models and feature subsets. However, the paper lacks a detailed explanation of the progressive sampling technique used, limiting the assessment of its effectiveness and potential limitations. It is worth noting that the auto-waikato environment for knowledge analysis (WEKA) approach [7] has a 30-hour time budget for each run, whereas the automatic model in this research averaged around 5 hours.

Ros and Guillaume [8], proposed an iterative sample selection method to enhance clustering performance. The authors utilize a fitness function to evaluate the relevance of each sample and handle the selection process. The paper evaluates the framework on benchmark datasets, comparing its performance to other clustering methods like k-means and hierarchical clustering. However, the paper lacks a detailed explanation of the fitness function used in the framework. Furthermore, the paper does not extensively compare the proposed framework with more recent and state-of-the-art clustering methods. Huang *et al.* [9], proposed a methodology to improve the accuracy of breast cancer prediction models. The paper highlights the challenges posed by imbalanced datasets in breast cancer prediction, where the minority class (cancer cases) has significantly fewer samples than the majority class (non-cancer cases). Effective feature selection and over-sampling techniques are emphasized to address this imbalance and enhance prediction performance. The proposed methodology combines the ReliefF feature selection algorithm with over-sampling techniques such as synthetic minority over-sampling technique (SMOTE) and Borderline-SMOTE. Despite the paper's contributions, there are some potential drawbacks. The ReliefF algorithm and its feature selection process are not thoroughly explained, limiting the understanding of its principles and effectiveness in breast cancer prediction. Zhao *et al.* [10], carried out research that suggests an approach for large-scale data clustering that is based on stratified sampling. The clustering analysis has utilized progressive sampling. According to the experimental findings, the suggested approach performs better than comparable algorithms for large-scale data sets in terms of clustering quality and computing efficiency. Zhang *et al.* [11], focuses on improving two-class unbalanced classification by investigating the effects of feature selection before and after data resampling. The decision between the two frameworks depends on variables such as the dataset, specific combinations of feature selection and data resampling techniques, and standard classification algorithms. The study also identifies the optimal combinations of feature selection and data resampling algorithms that consistently yield the best results in unbalanced classification.

Simulation research by Speiser [12] is conducted to compare the binary mixed model forest method (BiMM) with traditional generalized linear mixed model feature selection techniques like shrinkage and backward elimination. The study utilizes the health, ageing, and body composition study dataset to develop models for predicting mobility disability in older individuals and evaluates their performance. However, feature selection's impact on the performance of the binary mixed model (BiMM) forest method and its effectiveness is not adequately examined, highlighting a potential gap in understanding. Fei *et al.* [13], conducted a study on cotton classification at the county scale using Landsat 8 OLI images from the 2017 cotton growing season. They extracted spectral information, vegetation indices, and textural elements from the images as features for classification. The random forest feature selection method was applied to identify the most relevant features, and established machine learning techniques were employed for classification. The findings showed that combining multi-feature images improved accuracy and stability compared to single-phase images. While support vector machine (SVM) and artificial neural networks (ANN) outperformed random forest in classification results, random forest exhibited superior stability in multi-feature classification.

Xuan *et al.* [14], suggested a feature selection strategy using random forests and a hybrid neural network called convolutional neural network-bidirectional gated recurrent unit (CNN-BiGRU) for multi-model integrated prediction. It uses data from various sliding time windows of different lengths and trains separate sub-forecasting models. The predictions from all sub-models are combined to generate a single forecast. The input load dataset is resampled and sliding window widths of 5, 10, 20, and 50 are chosen. The method trains sub-forecasting models using different time frames and window widths. The model is trained for accurate

short-term load forecasting by averaging the projected results. Despite attempts to modify and optimize the power load data, the proposed approach struggles to accurately estimate ultra-short-term power loads. This suggests challenges in predicting power loads accurately within very short time intervals.

A PS transformer network for change detection in remote sensing images [15] was presented in this research. This network improves spatial connections and contextual information by iterative harvesting and optimizing feature information. The authors note that the a deeply supervised image fusion network for change detection (DSIFN-CD) dataset used in their study may have inaccurate omissions. The dataset's marker information identifies changing components by area rather than individual buildings, potentially overlooking important data and reducing irrelevant information. Furthermore, the dataset suffers from insufficient training data. However, the proposed PS-based system also faces the issue of over-extraction. This means the network may extract excessive or unnecessary information, leading to increased computational complexity or the inclusion of irrelevant features in the model.

Mahafzah *et al.* [16], presented the method that involved multiple stages of sampling, which potentially affected its effectiveness in real-time applications. The drawback of the proposed technique lies in its suitability for real-time applications. The multi-stage sampling process may require additional computational resources and time, making it less efficient for real-time data mining tasks that require immediate or near real-time processing and decision-making. Portet *et al.* [17], discovered that the progressive sampling approach could yield comparable results even when using only a third of the initial data. This indicates the effectiveness of the approach in reducing data size without significant quality loss. However, a drawback of the proposed method was increased time consumption. The progressive sampling process involved multiple phases, which extended the processing time compared to using the full dataset. Depending on specific application requirements and constraints, the benefits of reducing data size while maintaining satisfactory results may outweigh the additional processing time.

Villegas *et al.* [18], proposed an approach to enhance Twitter polarity detection through sentiment analysis. They suggested that by utilizing a specific subset of features, sentiment analysis models can be improved to achieve faster and simpler results. The authors employed rule-induction approaches to construct a reduced representation of tweets using the selected features, enabling the identification of relevant patterns and rules for sentiment analysis. They also utilized covering arrays to create bootstrap samples, aiding in estimating model performance and making predictions. However, the approach may not generalize well to sentiment analysis tasks beyond the Twitter platform. Zheng and Jin [19], introduced a method introduced for feature extraction from low sample size data. The method considers data quality and variable training samples. A key aspect of the proposed method is that the features are updated for each training sample. This dynamic adjustment of the feature selection process allows it to adapt based on the available training data. The authors likely conducted experiments to assess the performance of their method and compared it with other explicit feature selection techniques. The ability to update features for each training sample offers a more adaptive and flexible approach to feature selection in scenarios with a limited sample size.

Parthasarathy [20], used association rule mining and the idea of equivalence to achieve progressive sampling. The choice of feature importance algorithm along with the sampling method plays an important role for data classification. Chen *et al.* [21], used random forest as the features selection method to extract the best features along with random sampling. The results obtained are compared with various classifiers with and without feature extraction. Random forest performed better for feature extraction with various classifiers. The authors failed to explain about the processing time taken to train the model even though accuracy obtained is comparatively good. Sarada and Devi [22], utilized the under-sampling method combined with recursive feature elimination for best results but this method did not work well with some highly imbalanced datasets. The method presented by Hwang *et al.* [23], examined the near-infrared spectra's feature significance using random forest regression models. Better prediction performances were achieved by the random forest models trained with the high-importance regions than by those trained with the entire spectral range, proving the value of the feature importance calculated by the random forest and the viability of applying the spectral data selectively.

– Objectives

The main objectives of proposed research random forest regressor feature importance classification method based on progressive sampling (RFRFI-PS) are:

- 1) To develop an efficient classification model that can efficiently handle the time complexity of big data analytics.
- 2) To examine the efficacy of progressive sampling in increasing classification model efficiency and accuracy.
- 3) To assess the proposed model's performance using various classifiers and compare it to existing models.
- 4) To demonstrate the applicability of the proposed model on real-world datasets from different domains.

2. METHOD

The proposed RFRFI-PS model used “random forest regressor feature importance” technique for feature selection and employed progressive sampling to focus computational resources on informative data subsets. Through extensive experimentation, we aim to showcase the effectiveness of our model in reducing time complexity while maintaining high accuracy in big data analytics. Figure 2 depicts the flow design of the RFRFI-PS model.

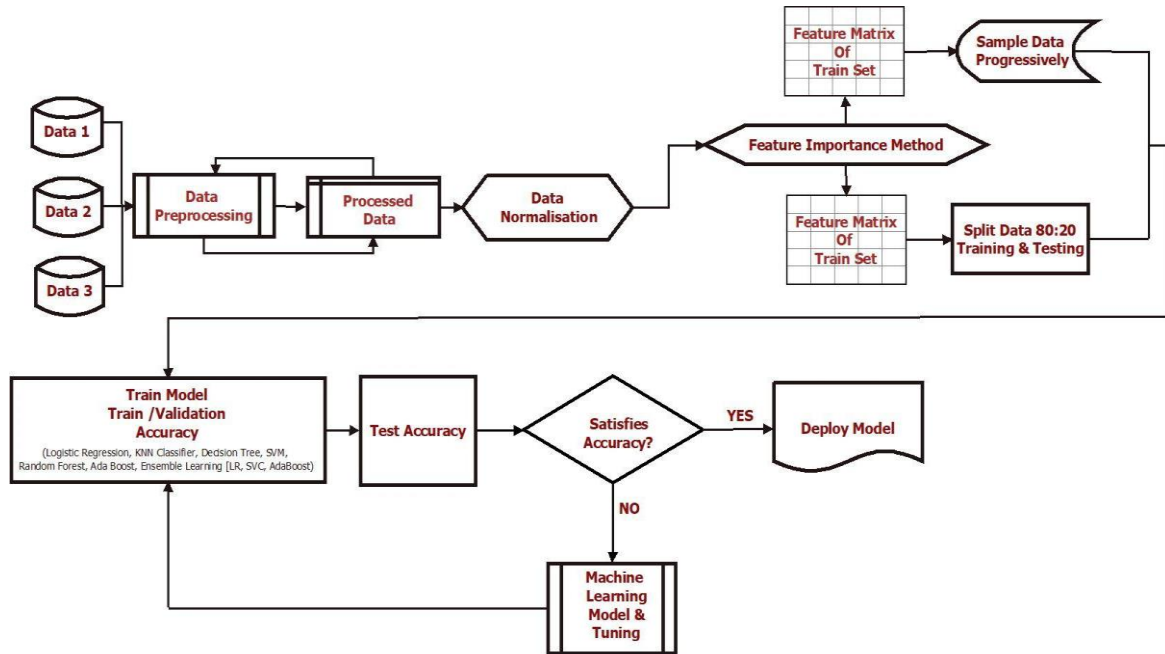


Figure 2. Block diagram of the proposed RFRFI-PS system

2.1. Dataset used for RFRFI-PS model

The performance evaluation utilized three datasets. The initial dataset under consideration is the cancer dataset. The provided dataset [24] pertains to the UCI machine learning repository, a repository that houses gene expression data of individuals diagnosed with various tumour kinds. The dataset aims to predict the type of cancer that an individual may acquire based on gene expressions.

The subsequent dataset [25] is referred to as the speech dataset. This dataset is associated with the Parkinson’s disease classification dataset included within the UCI repository. The dataset comprises recordings obtained from a sample of 64 individuals who were classified as healthy and served as the control group, as well as 188 individuals diagnosed with Parkinson’s disease, consisting of 81 women and 107 men.

The third dataset [26] under consideration is the human activity recognition with smart phones (HAR-SP) dataset available on Kaggle. The dataset comprises data that has been gathered from a cohort of 30 individuals who volunteered for the study, with ages ranging from 19 to 48 years. Each participant employed a mobile device to do six separate tasks, namely walking, ascending stairs, descending stairs, sitting, standing, and reclining.

2.2. RFRFI-PS model implementation

The proposed model involves four main steps: data preprocessing and normalization, and feature importance using the random forest regressor method. Here’s a breakdown of each step:

Step 1: data preprocessing: in this step, the standard approach to data preprocessing is followed. This typically involves handling missing values and outliers in the dataset. Null values and outliers are eliminated or treated in a way that does not negatively impact the analysis.

Step 2: normalization: after data preprocessing, a data normalization technique is applied to ensure consistent data values within a specific range or distribution. There are different normalization techniques available, such as min-max scaling or z-score normalization. The choice of technique depends on the dataset’s characteristics and analysis requirements. In our model building, we utilize min-max scaling normalization after considering the dataset’s characteristics and analysis requirements. Min-max

scaling, also known as normalization, rescales the data to a fixed range, typically between 0 and 1. The formula for min-max scaling is:

$$Z_SCALED = (Z - Z_MIN) / (Z_MAX - Z_MIN) \tag{1}$$

here, Z represents the original data, Z_MIN is the minimum value in the dataset, and Z_MAX is the maximum value. This method ensures that data values are proportionally mapped to the desired range while preserving the relationships between data points.

Step 3: feature importance using random forest regressor method: in this step, the importance of features in the dataset is determined. The impurity-based feature importance measure is employed, which utilizes the gini impurity criterion commonly used in decision tree algorithms. The gini impurity measures the effectiveness of a decision tree split in separating the data. Multiple decision trees are constructed using the random forest regressor, and the average impurity reduction for each feature is calculated. The relative values of the estimated importance indicate the significance of each feature. The feature with the highest decrease in impurity is selected for each internal node in the decision trees. This process is repeated for all features, and the average importance across all trees in the random forest is computed. This measure aids in identifying the most relevant features for the analysis.

The proposed method has been implemented by applying the feature importance method in step 2. Figure 3 shows the relative feature importance statistic. In Figure 3(a) to 3(c) explain about of cancer, HAR-SP and speech dataset respectively by applying the random forest regressor algorithm. The top features were selected. Initially, the top 10 features were extracted, followed by another 10. After conducting multiple rounds of trials, the top 80 features for the cancer dataset and the top 20 features for the HAR-SP and speech datasets were selected for the experimental process.

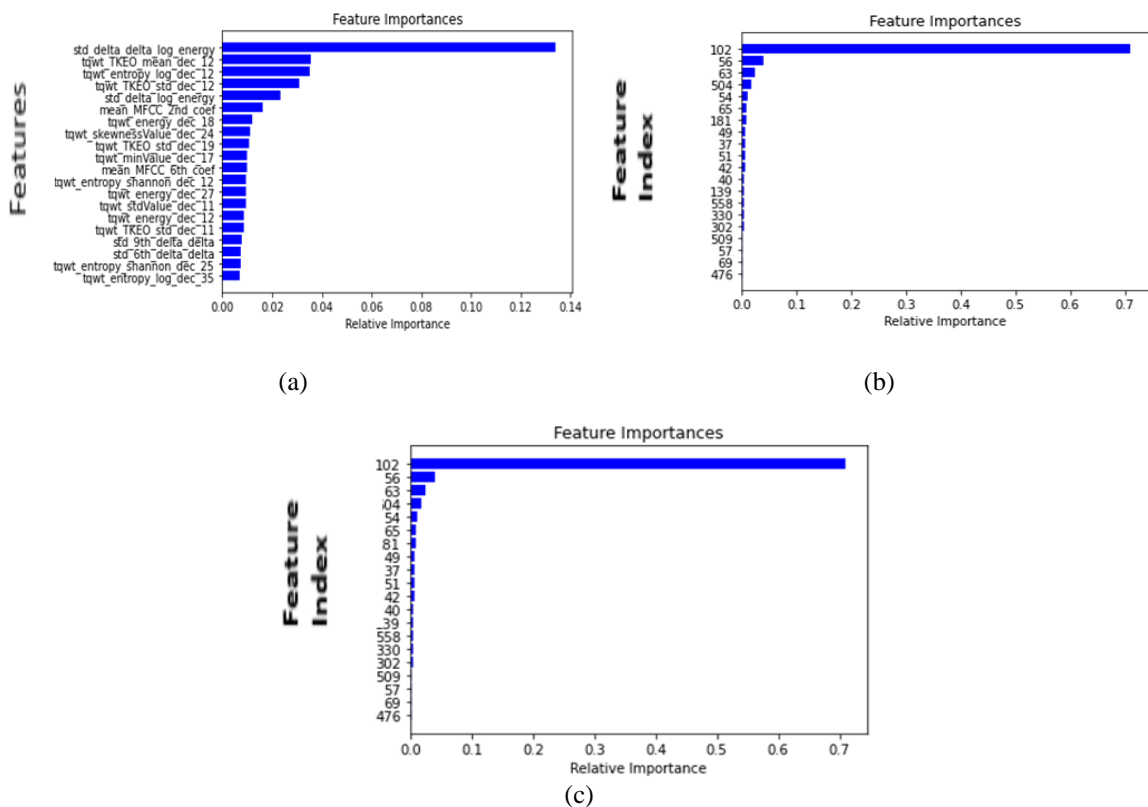


Figure 3. Important features of; (a) cancer dataset, (b) HAR-SP dataset, and (c) speech dataset

The result of applying the feature importance method is presented in Table 1, which provides information about the total number of reduced feature sets after selecting the best attributes for training the model. The Table 1 includes the count of feature instances (rows) and attributes (columns) before and after applying the feature importance method. The accuracy results obtained for each classifier are shown in Table 2.

Table 1. Characteristics of the original dataset and reduced dataset after selecting best feature attributes

Dataset		HAR-SP dataset	Speech dataset	Cancer dataset
Characteristics of the original dataset	Feature rows	4,252	756	800
	Feature columns	563	755	20,000
Reduced dataset after feature importance	Feature rows	4,252	756	800
	Feature columns	20	20	80

Step 4: step 4 of the proposed method is divided into two parts: 4a and 4b. Here’s a breakdown of each part:
 Step 4a: splitting the data into training and testing sets: in this part, the relevant feature data sample is divided into two sets: a training set and a testing set. The splitting is done in a ratio of 80:20, where 80% of the data is used for training the models, and 20% is reserved for testing and evaluating the accuracy of the models. Various classifiers are employed in this step to forecast the accuracy of the models.
 Step 4b: progressive sampling of the data: the progressive sampling approach starts with a small data sample i.e., 5% of the entire dataset, and gradually increases the sample size in multiples of 5%. The purpose of PS is to improve the accuracy of the model as quickly as possible by increasing the sample size used for training.

Table 2. Accuracy results in (%) for the reduced dataset for each classifier

Dataset	LR	KNN	SVM	DT	AdaBoost	RF	Ensemble
CANCER	99	93	97	86	83	91	98
SPEECH	86	84	80	84	80	84	85
HAR-SP	88	87	87	86	85	88	89

Progressive sampling intends to iteratively increase the sample size and monitor the model’s accuracy to determine if further increase in the sample size no longer leads to significant improvements in accuracy. The model stops once the curve attains the plateau. This approach reduces the classification time by considerable time. The accuracy of the models is evaluated using various classifiers, including logistic regression, k-nearest neighbors’ algorithm (KNN) classifier, SVM, decision tree, AdaBoost, ensemble learning, and random forest. The accuracy predictions for each sampling percentage and classifier are recorded and plotted in graphs. The graphs in Figures 4, 5, and 6 show the relationship between the percentage of data used and the accuracy achieved by each classifier for speech, human activity recognition with smartphones and cancer datasets.

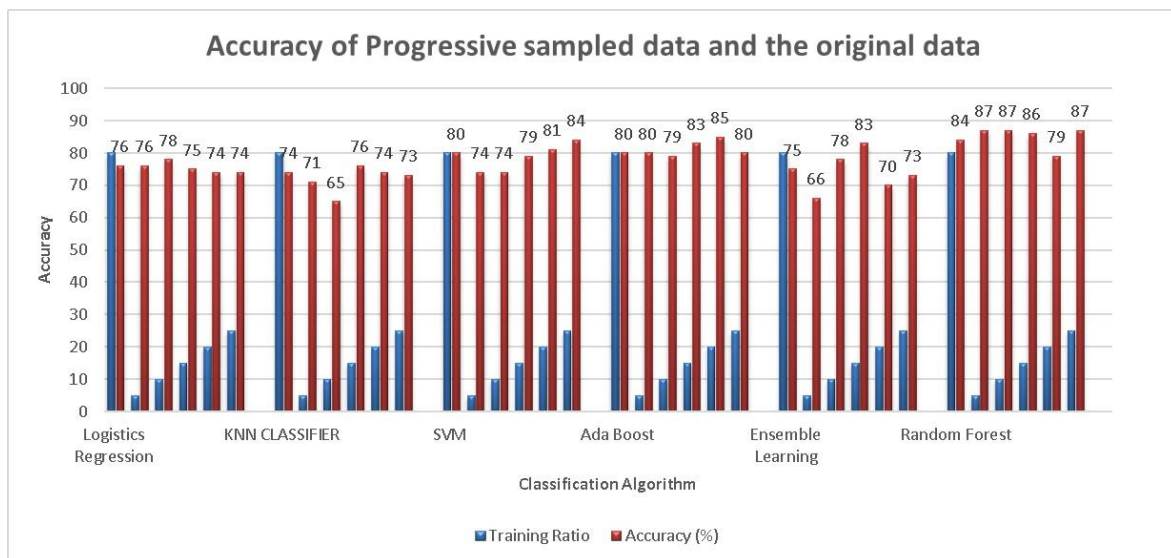


Figure 4. Visual representation sampled data v/s accuracy of speech data

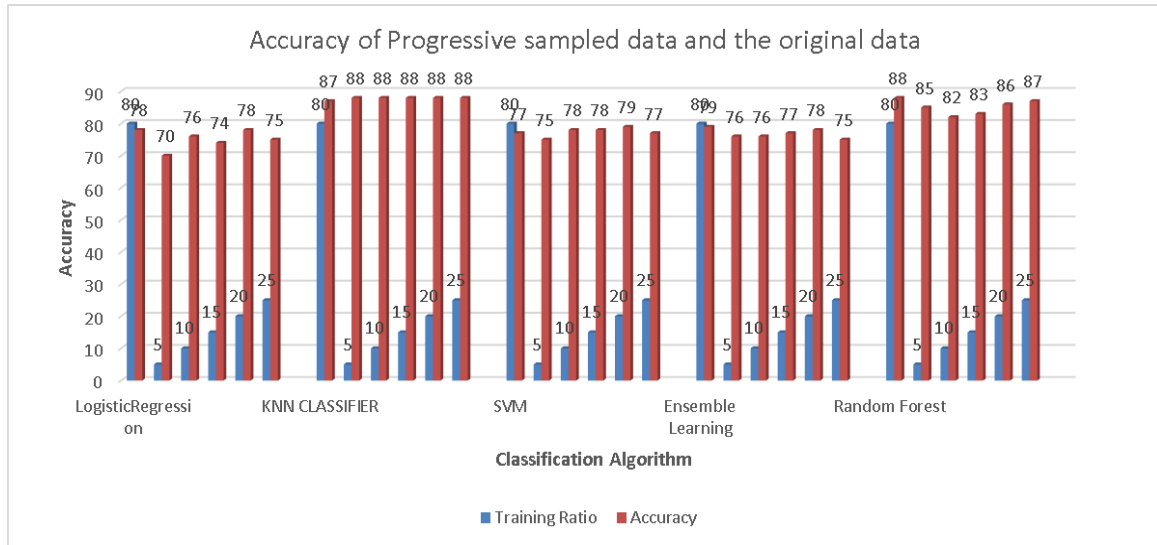


Figure 5. Visual representation sampled data v/s accuracy of HAR-SP data

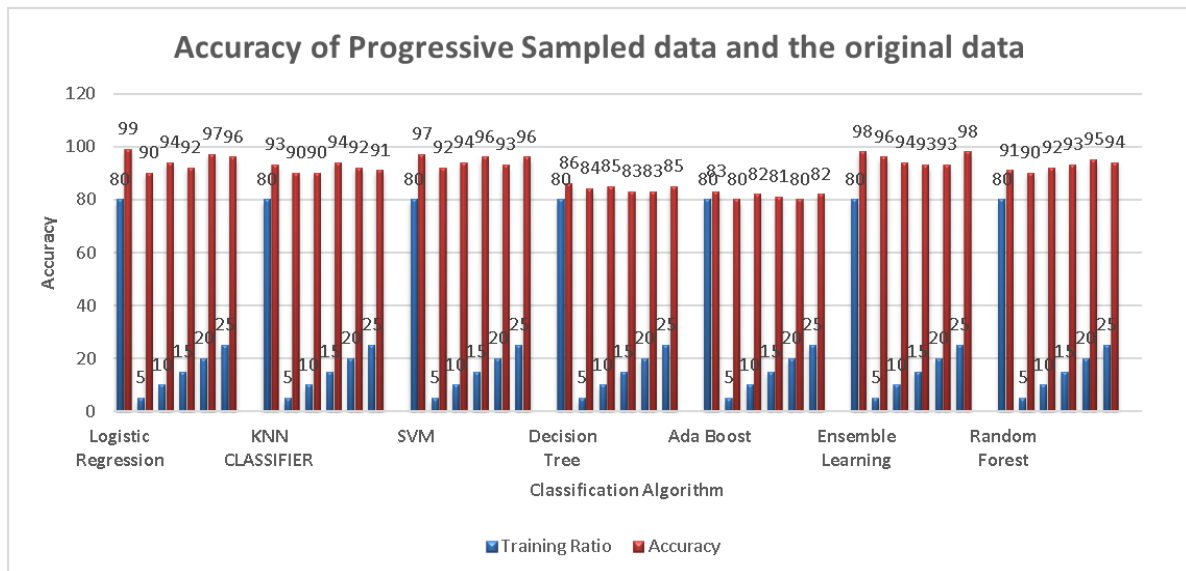


Figure 6. Visual representation sampled data v/s accuracy of cancer data

3. RESULTS AND ANALYSIS

The accuracy results of different classifiers at various sampling percentages provide insights into the model’s performance improvement as the sample size increases. It helps assess the trade-off between training data size and model accuracy. The computational efficiency of each classifier with different sampling proportions can be inferred from the timing information. In this proposed model, the classification efficiency was evaluated on cancer, speech, and HAR-SP datasets. The initial cancer dataset had 800 instances with 20,000 feature attributes, the speech dataset had 756 instances with 755 feature attributes, and the HAR-SP dataset had 4,252 instances with 563 feature attributes.

By applying random forest feature importance and preprocessing techniques, the datasets were reduced to 80 features for the cancer dataset and 20 features each for the speech and HAR-SP datasets. The relevant feature data sample was divided into 80% of the training data set and 20% of a testing dataset to evaluate model accuracy. Various classifiers were used, and logistic regression achieved accuracies of 98%, 86%, and 88% on the testing sets of the cancer, speech, and HAR-SP datasets respectively. Logistic regression successfully predicted class labels, demonstrating its effectiveness in classifying instances in these datasets. Additionally, an ensemble classifier achieved 89% accuracy in classifying the HAR-SP instances. To evaluate

the impact of progressive sampling, progressively sampled data ranging from 5% to 25% was tested by each classifier. The results were plotted in Figures 4, 5, and 6, showing the relationship between the percentage of data sampled and the accuracy achieved by different classifiers for each dataset. These graphs provide a visual representation of how the amount of sampled data influences classifier accuracy.

The proposed model compared the performance of classifiers using an 80:20 ratio split dataset and progressively sampled data. Unanticipatedly, even with a small sample, the accuracy achieved was similar to that of the 80:20 sample ratio data. Further testing with progressively sampled data ranging from 5% to 25% showed that accuracy remained within the same range. This suggests that classifiers can achieve comparable accuracy with smaller sample sizes, indicating the effectiveness of progressive sampling in reducing data requirements without compromising accuracy. Selectively sampling a smaller subset of the data reduces computational complexity and processing time while maintaining similar levels of accuracy.

Table 3 presents the execution time for the entire process, the whole dataset, and progressively sampled data representing different proportion. The results show that the execution time for progressively sampled data is notably lower compared to the entire dataset, but it can achieve approximately the same accuracy as the original dataset. Interestingly, despite the reduced sample size and time, the classifiers achieve similar accuracy on the sampled data compared to the entire dataset. These findings indicate that the proposed model can achieve comparable accuracy with fewer samples in a shorter time when working with datasets containing relevant features.

Table 3. Execution time of the whole dataset and progressive sampled data with different classifiers

Classifier	Cancer dataset (time in seconds)					
	SVM	KNN	Ada boost	Random forest	Ensemble	Logistic regression
80:20	8.6	12.69	3.40	5.44	10.5	7.31
5%	0.14	0.54	0.47	2.21	0.79	0.42
10%	0.30	0.68	0.77	2.23	1.02	0.66
15%	0.45	0.79	0.53	2.31	1.45	0.81
20%	0.73	1.14	0.71	0.9	1.91	1.22
25%	0.93	1.3	0.76	0.9	2.37	1.33
	Speech dataset (time in seconds)					
80:20	4.3	9.17	2.6	6.42	8.4	8.1
5%	0.3	3.20	0.4	0.21	0.4	0.8
10%	0.41	3.22	0.7	0.23	0.52	0.9
15%	0.43	3.91	0.6	0.30	0.62	1.12
20%	0.52	3.95	0.41	0.34	0.71	1.15
25%	0.54	3.83	0.31	0.44	0.83	1.23
	Human activity recognition with smartphones dataset (time in seconds)					
80:20	1.44	44	2.56	49	30	14.4
5%	0.15	4.6	0.49	6.12	1.5	4.9
10%	0.17	6.15	0.57	2.37	2.5	3.5
15%	0.42	8.19	0.73	3.18	4.0	2.2
20%	0.61	10.0	0.80	3.08	6.09	2.1

The utilization of progressive sampling and time complexity reduction techniques improves classification efficiency without sacrificing accuracy. The reduced execution time of sampled data is particularly beneficial in handling large datasets in big data scenarios, where processing time and computational resources are significant concerns. These results emphasize the model's efficiency in achieving accurate classification with reduced execution time, making it suitable for real-world applications involving big data, where both accuracy and efficiency are crucial factors to consider. The findings and outcome of the proposed system observed from the above results are as follows:

- The system effectively addresses data imbalance, heterogeneity, and multidimensionality by employing diverse classifiers (random forest, ensemble learning, SVM, KNN, and logistic regression) to handle different data types and achieve high accuracy in classification tasks.
- The proposed RFRFI-PS model introduces progressive sampling as a novel classification model, iteratively applying it to the dataset with the best features to significantly improve performance by optimally focusing computational resources on informative subsets, reducing time complexity.
- The system utilizes random forest regressor feature importance (RFRFI) with the gini significance method to identify important characteristics, effectively reducing data features for classification, and enhancing efficiency and accuracy.
- The system achieves over 85% accuracy with reduced data size (5-10% of the original), demonstrating accurate predictions while significantly reducing data processing requirements.
- The proposed system shows promising results in accuracy and time complexity reduction by combining

progressive sampling, feature selection using RFRFI, and diverse classifiers, effectively addressing challenges in large databases and improving classification accuracy.

3.1. Comparison of proposed RFRFI-PS model with research work

Table 4 compares the performance of the suggested approach with related work. In a previous study [27], logistic regression and random forest models were used for feature extraction and classification, but random sampling did not yield the expected results, as shown in the table. However, their training and testing process achieved good AUC scores.

Table 4. Comparison of proposed RFRFI-PS model with work

Research work	Dataset	Techniques	Machine learning algorithms	Performance evaluation metrics used and obtained results
Saarela and Jauhiainen [24]	Running injury data Breast cancer	Feature importance with random sampling using logistic regression	RF LR	AUC: 0.51+-0.06 AUC: 0.50+-0.02
Proposed RFRFI-PS model	Cancer, speech, HAR-SP	Feature importance with progressive sampling using random forest regressor	RF, LR, KNN, SVM, DT, ADA BOOST, ensemble learning	AUC: 0.51+-0.08 AUC: 0.50+-0.03 Average accuracy: 84.5%. Average F-measure with progressive sampling: 91% accuracy execution time: 1sec

In contrast, our suggested approach demonstrates high accuracy across various classifiers, especially on high-dimensional datasets. The incorporation of progressive sampling has significantly reduced the time required, which is particularly advantageous for large datasets. The cancer dataset, with its numerous feature dimensions, outperformed the human activity recognition with smartphones and speech datasets in terms of accuracy and F-measure. Hence, the progressive sampling method combined with feature extraction has proven valuable in reducing time in big data analytics.

3.2. Comparison of proposed RFRFI-PS model with research work

In a different study [28], three classifiers were employed on four high-dimensional DNA microarray datasets. Clustering was used to group related features based on manhattan distance. Random sampling was utilized for classification, and F-measure and accuracy were the performance metrics. It is important to note that our proposed work had a relatively larger number of instances compared to the dataset described in that study. Our work employed progressive sampling, which demonstrated accurate results in a few cycles. The random forest regressor approach was used for feature extraction, and the extracted features were classified using seven different classifiers. A comparison of the proposed model with work [28] can be found in Table 5.

Comparing our approach to the study [27], [28], our average F-measure of 91% is higher, indicating the effectiveness of our approach. The results highlight the superior performance of our proposed approach, with higher accuracy and F-measure. The use of progressive sampling, feature extraction, and a range of classifiers contributes to the success of our method in improving classification efficiency and reducing the time required in big data analytics shown in Table 3.

3.3. Inter-model performance comparison of progressive sampling methods using different techniques

In the domains of active learning and frequent mining, the progressive sampling approach was applied, although the specific domains differed from our study. Despite the domain differences, sampling was a common element across these fields. Our suggested strategy, which combines feature selection and progressive sampling, achieved an accuracy of 84.5%. In comparison, the accuracy obtained in the frequent mining and active learning domains was 78% and 79.9%, respectively. Detailed comparison results can be found in Table 6, highlighting the superior performance of our proposed method.

Table 5. Comparison of proposed RFRFI-PS model with work [28]

Research work	Dataset	Techniques	Classifier	Performance evaluation metrics-values
[28]	Colon, CNS, ovarian, lung	Feature importance method with spectral clustering and random sampling	MLP, SVM, C4.5	Average F-measure:83%
Proposed RFRFI-PS model	Cancer, speech, HAR-SP	Feature selection using random forest regressor feature importance combining progressive sampling	RF, LR, KNN, SVM, DT, ADA BOOST, ensemble learning	Average accuracy: 84.5%. Average F-measure with progressive sampling: 91% Execution time: 1 sec

The ramifications of the findings from the proposed system RFRFI-PS model are significant and have several potential implications for future research and practical applications:

- Improved efficiency in big data classification: the demonstrated use of progressive sampling and time complexity reduction techniques can significantly enhance the efficiency of big data classification tasks, becoming crucial for faster and more accurate data analysis amid the exponential growth of data volume.
- Scalable solutions for handling large datasets: the proposed system's high accuracy with reduced dataset size (5-10% of the original) provides a promising solution for handling large and complex datasets, ensuring scalability in data analysis without compromising accuracy.
- Generalizability across domains: the incorporation of diverse classifiers (random forest, ensemble learning, SVM, KNN, and logistic regression) and feature selection techniques like RFRFI renders the proposed system adaptable across diverse domains, offering a flexible approach for accurate classification of different data types in real-world scenarios.
- Resource optimization: the findings show that the proposed system optimizes time complexity by focusing on informative data subsets, offering broader implications for critical applications requiring computational efficiency and resource utilization.
- Real-world applications: the promising results in accuracy and efficiency are applicable across industries, benefiting real-world applications in healthcare, finance, marketing, IoT, enhancing decision-making, predictive modeling, and data-driven insights with large datasets.

Table 6. Comparative analysis of inter-model performance employing various techniques using progressive sampling method

Techniques used	Accuracy (%)
Active learning	79.9
Frequent mining	78.0
Proposed RFRFI-PS model (progressive sampling)	84.5 (average accuracy of all three datasets)

4. CONCLUSION AND FUTURE WORK

In our research, we combined progressive sampling and random forest regressor feature importance techniques to train clinical datasets. Our goal was to reduce high-dimensional datasets using feature importance. We utilized logistic regression, SVM, AdaBoost, ensemble, and random forest classifiers to evaluate accuracy and execution time on the reduced dataset. We progressively sampled the data from 5% to 25% and assessed accuracy and time using the same classifiers at each stage. The results showed that the progressively sampled data maintained an average accuracy of 85% (across all classifiers) while significantly reducing execution time to 1 second (across all classifiers) compared to 8 seconds for the entire dataset. This indicates the potential of our method in reducing time in big data analytics. The cancer dataset, with more features compared to the speech and human activity recognition with smartphones datasets, achieved higher accuracy. The effectiveness of the random forest regressor and its feature importance was evident in handling the extensive features and volume of the cancer dataset. Our proposed method offers a new direction in big data analytics, using progressive sampling and feature importance to reduce execution time while maintaining accuracy. These findings have implications for optimizing data analysis processes and advancing the field of data analytics.

For future research, it would be valuable to explore the combination of progressive sampling with diverse feature importance methodologies on datasets of varying sizes, class ratios, and feature counts. This could provide further insights into improving data analytics in terms of time and cost efficiency. The findings from this research can serve as a foundation for the development of more efficient and accurate classification systems in the era of big data. Researchers and practitioners may build upon the proposed system's methodologies, explore new combinations of classifiers and sampling techniques, and further optimize data processing in various application domains. The research also emphasizes the significance of balancing accuracy and computational efficiency, providing valuable insights for future studies aimed at improving the scalability and effectiveness of machine learning algorithms.





REFERENCES

- [1] C. Guo and J. Chen, "Big data analytics in healthcare," in *in Translational Systems Sciences*, 2023, pp. 27–70, doi: 10.1007/978-981-99-1075-5_2.
- [2] W. G. Cochran, "*Sampling Techniques*," Nashville, TN: John Wiley & Sons, 1977.
- [3] C. Davidson *et al.*, "Examining the effect of quality improvement initiatives on decreasing racial disparities in maternal morbidity," *BMJ Quality & Safety*, vol. 31, no. 9, pp. 670–678, Sep. 2022, doi: 10.1136/bmjqs-2021-014225.




- [4] H. Ahmad, B. Kasasbeh, B. Aldabaybah, and E. Rawashdeh, "Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS)," *International Journal of Information Technology (Singapore)*, vol. 15, no. 1, pp. 325–333, Jan. 2023, doi: 10.1007/s41870-022-00987-w.
- [5] V. Umarani and M. Punithavalli, "Analysis of the progressive sampling-based approach using real life datasets," *Open Computer Science*, vol. 1, no. 2, pp. 221–242, Jan. 2011, doi: 10.2478/s13537-011-0016-y.
- [6] X. Chen and S. Zeng, "Progressive sampling-based joint automatic model selection of machine learning and feature selection," in *Journal of Artificial Intelligence*, vol. 4, no. 1, pp. 30–38, 2021, doi: 10.23977/jaip.2020.040104.
- [7] V. W. Samawi, S. A. Yousif, and N. M. G. Al-Saidi, "Intrusion detection system: an automatic machine learning algorithms using auto-WEKA," in *2022 IEEE 13th Control and System Graduate Research Colloquium, ICSGRC 2022 - Conference Proceedings*, Jul. 2022, pp. 42–46, doi: 10.1109/ICSGRC55096.2022.9845166.
- [8] F. Ros and S. Guillaume, "A progressive sampling framework for clustering," *Neurocomputing*, vol. 450, pp. 48–60, Aug. 2021, doi: 10.1016/j.neucom.2021.04.029.
- [9] M. W. Huang, C. H. Chiu, C. F. Tsai, and W. C. Lin, "On combining feature selection and over-sampling techniques for breast cancer prediction," *Applied Sciences (Switzerland)*, vol. 11, no. 14, p. 6574, Jul. 2021, doi: 10.3390/app11146574.
- [10] X. Zhao, J. Liang, and C. Dang, "A stratified sampling based clustering algorithm for large-scale data," *Knowledge-Based Systems*, vol. 163, pp. 416–428, Jan. 2019, doi: 10.1016/j.knosys.2018.09.007.
- [11] C. Zhang *et al.*, "An empirical study on the joint impact of feature selection and data resampling on imbalance classification," *Applied Intelligence*, vol. 53, no. 5, pp. 5449–5461, Jun. 2023, doi: 10.1007/s10489-022-03772-1.
- [12] J. L. Speiser, "A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data," *Journal of Biomedical Informatics*, vol. 117, p. 103763, May 2021, doi: 10.1016/j.jbi.2021.103763.
- [13] H. Fei *et al.*, "Cotton classification method at the county scale based on multi-features and random forest feature selection algorithm and classifier," *Remote Sensing*, vol. 14, no. 4, p. 829, Feb. 2022, doi: 10.3390/rs14040829.
- [14] Y. Xuan *et al.*, "Multi-model fusion short-term load forecasting based on random forest feature selection and hybrid neural network," *IEEE Access*, vol. 9, pp. 69002–69009, 2021, doi: 10.1109/ACCESS.2021.3051337.
- [15] X. Song, Z. Hua, and J. Li, "PSTNet: progressive sampling transformer network for remote sensing image change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8442–8455, 2022, doi: 10.1109/JSTARS.2022.3204191.
- [16] B. A. Mahafzah, A. F. Al-Badarnah, and M. Z. Zakaria, "A new sampling technique for association rule mining," *Journal of Information Science*, vol. 35, no. 3, pp. 358–376, 2009, doi: 10.1177/0165551508100382.
- [17] F. Portet, F. Gao, J. Hunter, and R. Quiniou, "Reduction of large training set by guided progressive sampling: application to neonatal intensive care data," in *Intelligent Data International Workshop on Analysis in Medicine and Pharmacology (IDAMAP2007)*, Amsterdam, Netherlands, pp. 1–2, 2007.
- [18] J. Villegas, C. Cobos, M. Mendoza, and E. Herrera-Viedma, "Feature selection using sampling with replacement, covering arrays and rule-induction techniques to aid polarity detection in twitter sentiment analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11238 LNAI, 2018, pp. 467–480.
- [19] W. Zheng and M. Jin, "Improving the performance of feature selection methods with low-sample-size data," *Computer Journal*, vol. 66, no. 7, pp. 1664–1686, Jul. 2023, doi: 10.1093/comjnl/bxac033.
- [20] S. Parthasarathy, "Efficient progressive sampling for association rules," in *Proceedings - IEEE International Conference on Data Mining, ICDM, 2002*, pp. 354–361, doi: 10.1109/icdm.2002.1183923.
- [21] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 1, p. 52, Dec. 2020, doi: 10.1186/s40537-020-00327-4.
- [22] C. Sarada and M. S. Devi, "Imbalanced big data classification using feature selection under-sampling," *CVR Journal of Science & Technology*, vol. 17, no. 1, pp. 78–82, Dec. 2019, doi: 10.32377/cvrjst1714.
- [23] S. W. Hwang *et al.*, "Feature importance measures from random forest regressor using near-infrared spectra for predicting carbonization characteristics of kraft lignin-derived hydrochar," *Journal of Wood Science*, vol. 69, no. 1, p. 1, Jan. 2023, doi: 10.1186/s10086-022-02073-y.
- [24] Gene expression cancer RNA-Seq, UCI Machine Learning Repository, 2016, doi: 10.24432/C5R88H.
- [25] Parkinsons, UCI Machine Learning Repository, 2008, doi: 10.24432/C59C74.
- [26] Human Activity Recognition Using Smartphones, UCI Machine Learning Repository, 2012, doi: 10.24432/C54S4K.
- [27] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences*, vol. 3, no. 2, p. 272, Feb. 2021, doi: 10.1007/s42452-021-04148-9.
- [28] S. Liu and K. Zhang, "Under-sampling and feature selection algorithms for S2SMLP," *IEEE Access*, vol. 8, pp. 191803–191814, 2020, doi: 10.1109/ACCESS.2020.3032520.

BIOGRAPHIES OF AUTHORS






Nandita Bangera     has completed her B.E. in Computer Science and M.Tech. in Software Engineering and currently pursuing her Ph.D. She has 15 years of experience in the field of teaching and currently, she is pursuing her research in big data analytics. Has published 2 papers in international journal and 3 national journals. She has actively participated in and coordinated various seminars, conferences workshops, and FDPs. Have guided students in UG as well PG level projects and have received grants for the best final year project. Has coordinated training programs conducted by Infosys to make student's placement ready. Her area of interest is web technology and machine learning. She can be contacted at email: nanditamanohar@gmail.com.






Dr. Kayarvizhy    completed her Bachelors's and Master of Technology in Computer Science from Pondicherry University. She holds a Ph.D. degree from Anna University, Chennai. She has over 17 years of teaching experience. Her research interests include machine learning, data science, the internet of things, and software metrics. She has 30 publications in various international journals and conferences. She continues to foster newer innovations through government grants like VGST, Dassault systems, and KSCST in her field of interest. She can be contacted at email: kayarvizhyn.cse@bmsce.ac.in.



Shubham Luharuka    is a Machine Learning Engineer at CloudSEK Information Security Pvt. Ltd, located in Bangalore, India. He specializes in the diverse application of natural language processing (NLP) and computer vision in the field of cybersecurity. He is also an instructor for the NVIDIA Deep Learning Institute and a Google Crowdsourcing ML Facilitator. He has authored several papers including IEEE. He has pursued extensive learning in machine learning, artificial intelligence, and quantum machine learning. He has completed numerous online certification courses from esteemed organizations such as NPTEL, Stanford University, and etc. He gained valuable experience as an AI Software Intern at Synapsica Healthcare Pvt. Ltd., as well as a Data Science Intern at Exposys Data Labs and Informatics Research Lab. He actively participates in hackathons and other challenges to further enhance his skills and knowledge. His keen interest in parallel programming, coupled with his passion for utilizing machine learning and deep learning, motivates him to explore their applications in diverse domains such as healthcare, agriculture, security, and space. He can be contacted at email: shubhamluharuka.cs.23@gmail.com.



Dr. Asha S Manek    is currently working in the Department of Computer Science and Engineering, Professor T. John Institute of Technology Bangalore, India. She received her B.E. degree from Nagpur University in 1993 and her M.E. in Information Technology from the University Visvesvaraya College of Engineering, Bangalore University, Bangalore in 2008. She holds a Ph.D. Degree in Computer Science and Engineering from JNTU, Hyderabad, India in 2019. In her 28 years of teaching experience at both UG and PG levels, she has published more than 45 research papers in various national and international conferences and journals including in Scopus indexed, SCI publications, IEEE Access, Springer, WoS, and Inderscience. Her research interests include data mining, machine learning, web mining, information retrieval, soft computing, text mining, social media analysis, IoT, cloud computing, and wireless sensor networks. She worked as a reviewer for various journals and conferences. She was honored as Judge, Resource Person, Session chair, TPC member, and Advisory committee member at National and International Seminars, Workshops, Conferences, Project Competitions, Hackathon, Kavach, and Toyathon. She is also an IEEE, CSI member, and a lifetime member of ISTE. She can be contacted at email: ashasm.cse@gmail.com.