IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 November 2022, accepted 6 December 2022, date of publication 26 December 2022, date of current version 6 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232404

RESEARCH ARTICLE

Classification of Polluted Silicone Rubber Insulators by Using LIBS Assisted Machine Learning Techniques

K. SANJANA[1], MYNENI SUKESH BABU[2], RAMANUJAM SARATHI[2], (Senior Member, IEEE), AND NARESH CHILLU[3], (Member, IEEE)

[1]School of Computing and Electrical Engineering, IIT Mandi, Mandi 175001, India
[2]Department of Electrical Engineering, IIT Madras, Chennai 600036, India
[3]School of Electronics Engineering, VIT-AP University, Amaravati, Andhra Pradesh 522237, India

Corresponding author: Naresh Chillu (naresh.ch@vitap.ac.in)

ABSTRACT Silicone rubber (SR) samples are coated with various types of artificially prepared pollutants, in order to identify and distinguish them by employing laser induced breakdown spectroscopy (LIBS). LIBS analysis is successful in identifying the elemental composition of the various types of pollutants. The presence of copper sulphate as well as carbon-based compounds such as fly ash, coal and calcium-based compounds such as cement and calcium phosphate (fertilizer) have been identified by the increment in the normalized intensity ratio of the copper, carbon and calcium peaks respectively. LIBS spectral data has been used in conjunction with several machine learning (ML) algorithms such as linear discriminant analysis, decision tree, K-nearest neighbors and various gradient boosting techniques to classify seven different types of contaminated SR samples. When compared to the other ML approaches utilized in the present study, classification using the Light gradient boosting technique has reflected better classification accuracy of 97.43% with a reasonable computation time of 5.1 s.

INDEX TERMS Gradient boosting, insulator, LIBS, machine learning, pollution, silicone rubber.

I. INTRODUCTION

With their promising features over the traditional insulators, the polymeric insulators are becoming increasingly common in power system network. Due to their outstanding qualities such as low weight, better vandal resistance, good surface hydrophobicity and high mechanical strength, the silicone rubber insulators are widely getting popular in high voltage outdoor insulation [1], [4]. Formation of water droplets on the insulating material surface owing to environmental conditions such as rain, fog, or snow as well as temperature gradient and salt deposition on the insulator surface due to surrounding ambience, are the primary factors that have influence on the behaviour of outdoor polymeric insulators [5], [7]. The outdoor polymeric insulators are also vulnerable to contamination from human activities such as fertilizer deposition and industrial activities such as smoke and chemical depositions [8], [9]. Pollution layer that builds up on the insulator surface, after absorbing moisture leads to a conductive path formation for a substantial leakage current to flow [10], [11]. Deposition of pollutants such as carbon, NaCl, gypsum, KNO_3, along with metal contaminations such as Ni and Cu will tend to cause frequent surface flashovers on the insulator surface [8], [12], [13]. The impact of aluminium phosphate fertilizer, as a contaminant on the surface of insulators near to fields, has been investigated recently [14], [15]. The hydrophobicity and surface flashover voltage of SR insulators are diminished drastically, when they are contaminated. Water film formation becomes easier on the SR surface, when its hydrophobicity is decreased, resulting in the development of the path for leakage current propagation [11]. Also, it is indicated that the mixture of NaCl with other salts, as an artificial pollutant had a significant impact on the surface flashover voltages [7]. According to the literature, it is important to identify and categorize the various contaminants that are dispersed throughout the surface of the insulators according to the level of pollution.

The associate editor coordinating the review of this manuscript and approving it for publication was Rajesh Kumar.

Laser induced breakdown spectroscopy (LIBS) is regarded as one of the promising techniques for elemental analysis of various materials [16], [18]. When compared to the conventional techniques, the LIBS technique has the significant advantage of identifying the pollutant deposit present on the surface fo the insulator without interrupting the power supply, with minimal (almost zero) damage to the insulator surface. Further, due to its benefits, like rapid and remote measurement capability, the popularity of the LIBS approach has been widely increasing in various fields [19], [20].

Using LIBS, Wang et al. found several pollutants that are deposited on the high voltage transmission line insulators [21]. Kumar et al. used the LIBS approach to determine the concentration of salt-type contaminants deposited on the surface of the wind turbine blades [22]. Also, the LIBS technique was employed for the analysis of various types of coal samples [23], [24]. Recently, Remote LIBS analysis has been employed for the identification of salt-deposition on the 33 kV SR insulator from a distance of 15 m [25]. As a result, the LIBS method has been used in this study to determine the elemental characterization of several types of pollutants deposited on the surface of the SR insulator.

Machine learning (ML) techniques are well-established as prominent tools for classification and prediction (regression) of various types of data. LIBS assisted with these ML techniques like support vector machine (SVM), Random Forest (RF), principal component analysis (PCA) and artificial neural network (ANN) are gaining popularity in material classification [26], [28]. Costa et al. classified polymer e-waste using KNN and soft independent modeling of class analogy (SIMCA) and obtained average classification accuracies as 98 percent and 92 percent respectively [29]. Lin et al. integrated LIBS with machine learning techniques such as support vector machine (SVM) and Boosting techniques along with optimization using PCA as well as RF methods, to distinguish boundary tissues and lung tumor tissues and found that boosting tree method with RF yielded an accuracy of 98.9 percent [30]. Gaudiuso et al. have used classification techniques such as linear discriminant analysis (LDA), Fischer's discriminant analysis (FDA), SVM and Gradient boosting to detect melanoma for early diagnosis and indicated that the classification accuracy of gradient boosting was higher with 97% [31]. In the present study, by considering the above literature, an effort has been made to classify the LIBS data with machine learning techniques like LDA, decision tree method, KNN and different boosting techniques.

Having known all these, the present work is concentrated, (a) to perform elemental analysis of polluted SR insulators for identifying the composition of pollutant and (b) classification of polluted SR insulators by employing various machine learning techniques such as linear discriminant analysis (LDA), Decision tree method, K-nearest neighbors (KNN) technique, boosting methods like gradient boosting, histogram-based boosting and advanced gradient boosting techniques such as extreme gradient boosting method and light gradient boosting method, to LIBS analysis. In addition, a comparison of different ML approaches in terms of classification accuracy of training, testing and overall data has been presented in the current work. Fig. 1 represents the framework of the process carried in the present work.

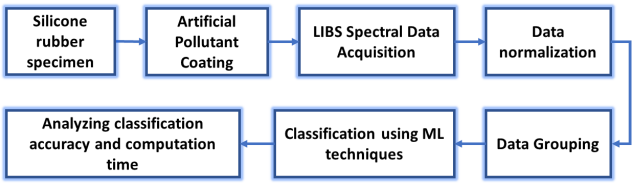

FIGURE 1. Framework of the process carried in the present work.

II. SAMPLE DETAILS AND EXPERIMENTAL SETUP

A. SAMPLE DETAILS AND CONTAMINATION PROCEDURE

A portion of 33 kV silicone rubber insulator has been considered for performing experimental investigations in the present study. Artificial pollutant slurries prepared with seven distinct types of pollutant mixtures (shown in Table 1) have been applied uniformly on the surface of the SR insulator. The kaolin clay is the composition of 30-40% of Al_2O_3, 0-3.2% of Fe_2O_3, 40-50% of SiO_2 and 7-14% of H_2O [32], which is mixed along with pollutants and deionized water in order to form an artificial pollutant slurry. The weight of NaCl and $CuSO_4$ are selected such that the mass fraction of Na and Cu will be 14.28 % and 3.65 % respectively [32]. These mass fractions of Na and Cu represent "very high" level of pollutants, as per IEC/TR2 61245:1993 [13]. The percentage weight of remaining pollutants such as fly ash, coal, cement and Ca_3PO_4(fertilizer) is maintained at 3%.

TABLE 1. Type of pollutant and its composition.

Pollutant type	Composition
Type-1	Kaolin clay + NaCl
Type-2	Kaolin clay + NaCl + $CuSO_4$
Type-3	Kaolin clay + NaCl + $CuSO_4$ + Fly ash
Type-4	Kaolin clay + NaCl + $CuSO_4$ + Coal
Type-5	Kaolin clay + NaCl + $CuSO_4$ + Cement
Type-6	Kaolin clay + NaCl + $CuSO_4$ + Ca_3PO_4 (Fertilizer)
Type-7	Kaolin clay + NaCl + $CuSO_4$ + Ca_3PO_4 + Coal

B. LIBS SETUP

The Nd3+: YAG laser source of model number: LAB-150-10-S2K, manufactured by Quanta-Ray LAB series, Spectra Physics, France has been used in the LIBS experimental setup to generate a laser beam with a wavelength of 1064 nm and a pulse width of 10 ns (Fig. 2). Through the focusing lens (25 cm focal length), the laser beam is directed at the target. The optical emission coming from the target after laser ablation is focused using a lens (100 cm focal length) and

collected using signal collector which is connected to a spectrometer through an optical fiber. In the current study, the spectral data was analyzed between 200 and 800 nm. The laser pulse energy in the current study was fixed at 40 mJ. The generated laser beam has a spot diameter of 0.5 mm. The optical fiber that is used to capture the emission occurred during the plasma formation has a core diameter 400 μm, 0.22 numerical aperture which is connected to the spectrometer (Ocean Optics USB2000+ UV–VIS-ES). The optical resolution of the spectrometer was 0.3 nm full width half maximum. The further details of the instrument specifications employed in the experimental setup are provided in our previous study [32].

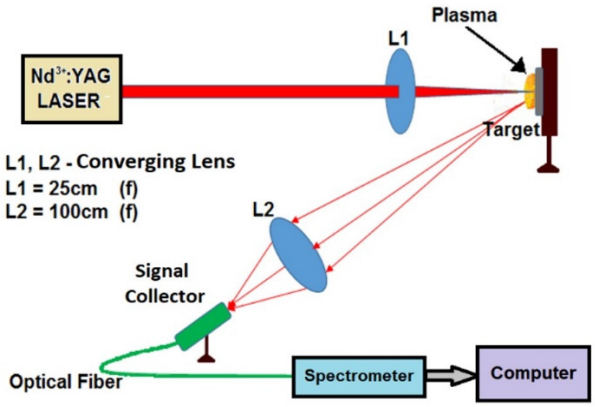

FIGURE 2. Experimental setup for Laser-Induced Breakdown Spectroscopy.

III. RESULTS AND DISCUSSION

A. ELEMENTAL ANALYSIS USING LIBS SPECTRA

The LIBS spectra of the polluted SR specimens are depicted in Fig. 3 and 4. Using NIST database, spectral peaks corresponding to Na, C, O, Si, N, Al, Ca and Cu were identified on the surface of the test samples [33]. From Fig 3 and 4, the presence of Na I peak (588.9 nm) indicates the deposition of NaCl on its surface. The Cu I (324.7 nm) spectral peak from Fig 3 and 4 indicates the presence of copper as contaminant on the SR surface. In case of Type-1 pollutant (as depicted in Fig 3a), where the $CuSO_4$ is not added, no significant peak at 324.7 nm has been noticed. Since, the fly ash consists of carbon as well as calcium compounds, spectral peaks corresponding to Ca II (396.8 nm) and C II (657.8 nm) have been noticed in Type-3 pollutant deposited SR specimens.

A significant increment in the C II (657.8 nm) indicates the presence of the coal in the Type-4 pollutants. In case of Type-5 and Type-6 pollutants along with Na I peak (588.9 nm) and Cu I (324.7 nm) spectral peaks, significant increment in the Ca II (396.8 nm) peak have been noticed. It is because, along with NaCl and CuSO4, the Type-5 and Type-6 pollutants consists of calcium compounds such as cement and Ca3PO4 respectively. Type-7 pollutant, which is a combination of all the compounds such as NaCl, CuSO4, Ca3PO4 and coal has reflected the significant increase in Na I peak (588.9 nm), Cu I (324.7 nm), Ca II (396.8 nm) and C II (657.8 nm) peaks respectively.

It is well-established that the normalized intensity ratio can substantially identify the variations in the quantity of an element present in the material [18], [32]. Hence, the normalized intensity ratios of the spectral peaks shown in the Fig 3 and 4 are calculated by taking neutral sodium (Na I) peak (588.9 nm) as reference in all the cases. It is noticed that the normalized intensity ratios of Ca II (396.8 nm) in case of Type-3, 5, 6 and 7 are 0.32 ± 0.02, 0.34 ± 0.04, 0.66 ± 0.04 and 0.58 ± 0.05 respectively. This significantly higher than the normalized intensity ratio of Ca II (396.8 nm) peak in case of Type-1 pollutant, where there are no calcium compounds. Therefore, the presence of calcium compounds

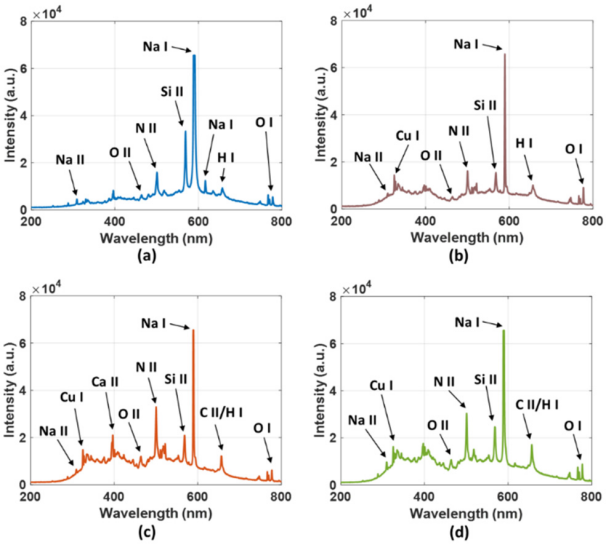

FIGURE 3. LIBS spectra corresponding to SR insulator coated with (a) Type-1, (b) Type-2, (c) Type-3 and (d) Type-4 pollutants.

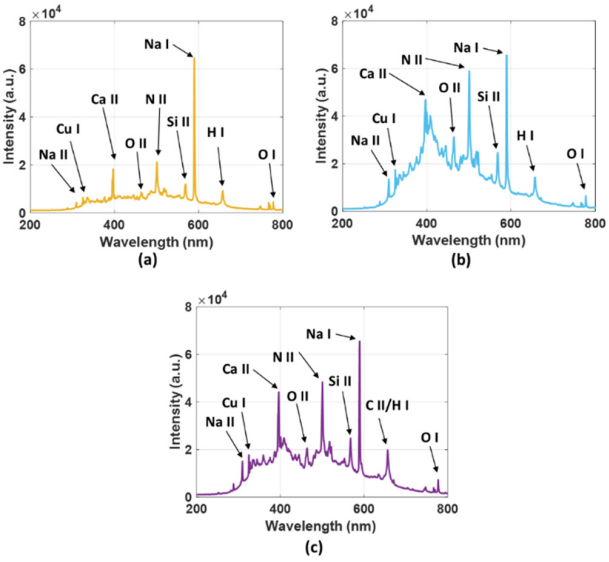

FIGURE 4. LIBS spectra corresponding to SR insulator coated with (a) Type-5, (b) Type-6 and (c) Type-7 pollutants.

in these types of pollutants has been reflected by the normalized intensity ratio of Ca II (396.8 nm) spectral peak.

Also, the normalized intensity ratios of C II (657.8 nm) in case of Type-3, 4 and 7 are 0.19 ± 0.02, 0.27 ± 0.03 and 0.31 ± 0.03 respectively. These are significantly higher compared to the normalized intensity ratios of C II (657.8 nm) in case of Type-1 pollutant, indicating the presence of the carbonaceous compounds. Hence, the LIBS technique can be adopted for identifying the type of pollution depositions on the insulators by determining the normalized intensities ratio. For a better identification of these mixture of pollutants and to categorize these polluted SR insulators, various ML techniques have been employed to the LIBS spectral data.

B. CATEGORIZATION OF POLLUTED SR INSULATORS USING ML TECHNIQUES

For distinguishing different types of contaminants present on the SR surface, various machine learning techniques such as linear discriminant analysis (LDA), decision tree method, K-nearest neighbors (KNN) technique, boosting methods such as gradient boosting, histogram-based boosting and advanced gradient boosting techniques such as extreme gradient boosting method and light gradient boosting method, have been employed to LIBS spectral data of seven different mixtures of pollutants (as shown in Table-1). As input dataset, a total of 1400 observations of seven different types were selected, with 200 observations belonging to each type of pollutant. The total number of features collected in the given dataset is 2048.Thus, the input dataset will have the dimensions of 1400 × 2048 and it has been divided into training and testing dataset in the percentage of 80 % and 20% respectively.

1) LINEAR DISCRIMINANT ANALYSIS (LDA)

LDA is mainly employed for dimension reduction and classification, which projects the dataset from its original dimension of features to a reduced dimensional space. In this method, each class is fitted according to Gaussian distribution and it is considered that each class has the same covariance matrix. The decision boundary generated, which is linear, is a product of fitting class conditional densities to the data set. Here, the dataset is classified with respect to each class with the likelihood amongst all classes using Bayes rule. The posterior probability is calculated according to Bayes rule as shown in (1).

$$P(y = k|X = x) = \frac{pi_k f_k\,(x)}{\sum_{l=1}^{K} pi_k f_k\,(x)} \tag{1}$$

where, pi_k is the prior probability of class k and $f_k(x)$ is the probability density function. After fitting Gaussian density on the above equation, we obtain the discriminant function as depicted in (2).

$$\delta_i(x) = log(f_k(x)) + log(pi_k) \tag{2}$$

where, the discriminant function $\delta_i(x)$ indicates the likelihood of data instance x for each class. Then, the decision boundary is the set of all the data instances whose discriminant function value is the same.

In the current classification problem, we use the default solver parameter, *svd* (singular value decomposition). Using this solver parameter is advantageous in cases where the number of features is large since it does not depend on calculation of covariance matrix. Absolute threshold parameter was set to 10^{-4}. LDA_Classifier function in Python with the set parameters has been used to classify the current dataset. Fig. 5 depicts the confusion matrix of polluted SR samples classified by using LDA classifier. Using this ML technique, the classification accuracy of training and testing dataset are obtained as 100% and 74.29% respectively. Thus, an overall classification accuracy of 94.86% was obtained using this method (Fig. 5). The program execution time for LDA classification model is noticed to be 2.2s.

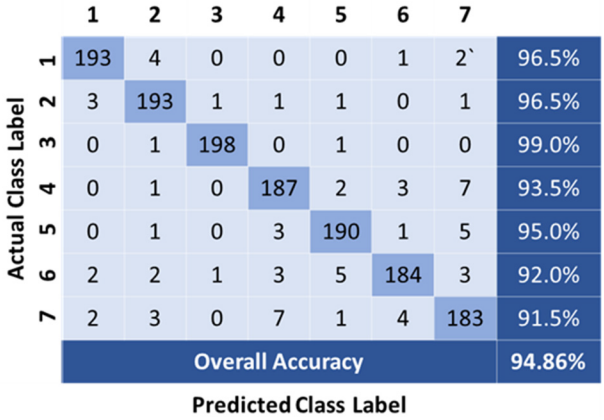

FIGURE 5. Confusion matrix determined by using LDA classifier.

2) DECISION TREE METHOD

A decision tree is basically a tree-like structure which consists of three main elements such as root node, branches and leaf node. Root Node represents the top most decision node. Branches represents the chance outcome. Leaf Node: represents the decision i.e., the final result. The tree is learnt by splitting the dataset into smaller subsets, which is performed by an attribute value test. This action is repeated several times iteratively. The classification of the data starts at the root node, where the attribute specified by that particular node is tested.

Subsequently, the branch corresponding to the attribute value obtained, is moved down and same steps mentioned above on the subtree are performed. DecisionTreeClassifier from *sklearn.tree* Python module with the set parameters has been adopted to classy the present dataset. The parameters were set to its default value with n_estimators as 100, the function to measure quality of split as *entropy* and min_samples_split as 2. Fig. 6 depicts the confusion matrix of polluted SR samples classified by using decision tree method. Using this technique, an overall classification accuracy of 94.86% was obtained, with training and testing accuracy as 100% and 74.29% respectively. The total

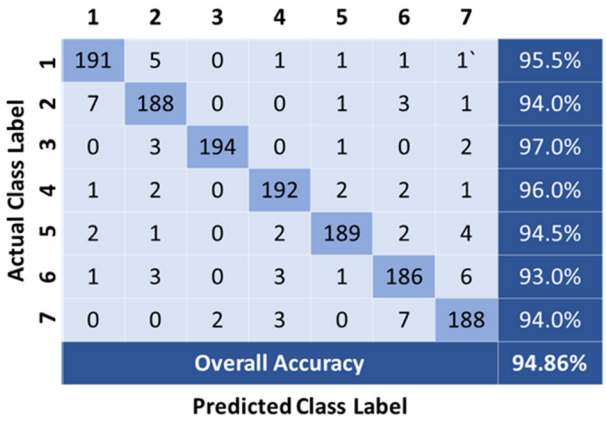

FIGURE 6. Confusion matrix determined by using decision tree classification method.

program execution time for the decision tree classification model is obtained as 4.0s.

3) K-NEAREST NEIGHBORS (KNN)

KNN is a neighbors-based classification technique, which is one of the most commonly used supervised learning algorithms. In this method, it is assumed that the data instances with similar features exist in close proximity. The classification is performed by calculating the majority vote of the nearest neighbors of each data instance. The Euclidean distance d, which is used to determine the proximity is calculated as shown in (3)

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

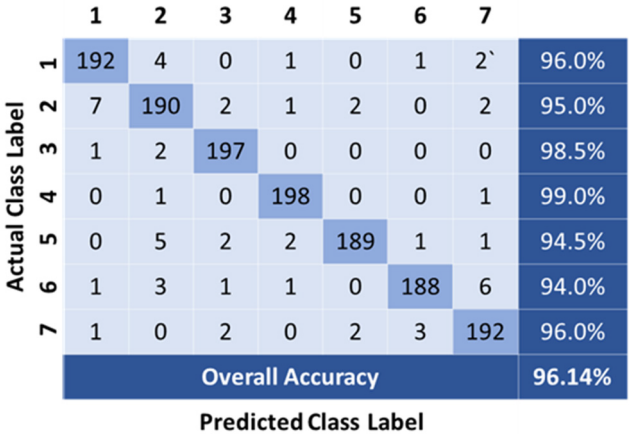

FIGURE 7. Confusion matrix determined by using KNN classification method.

Here, the value of k is initialized to choose the number of neighbors. For each data instance, the Euclidean distance is calculated. Then the first k nearest neighbors are chosen after sorting the above calculated distances. The number of data instances corresponding to each class are counted and finally, the data instance is assigned to the class with the maximum number of neighbors. Then, the parameter weights are assigned such that they are proportional to the inverse of distance calculated. The weights are implemented to ensure that the nearer neighbors contribute more to the fit. *sklearn.neighbors* Python module with the set parameters has been adopted to classy the present dataset. Fig. 7 depicts the confusion matrix of polluted SR samples classified by using KNN classifier. Using this ML technique, the classification accuracy of training and testing dataset are obtained as 100% and 80.71% respectively. Thus, an overall classification accuracy of 96.14% was obtained using this method (Fig. 7). A program execution time of 0.6s has been taken for classifying the polluted SR samples using KNN classifier.

4) GRADIENT BOOSTING METHOD (GBM)

Gradient Boosting is an ensemble learning technique which performs well with both heterogeneous and smaller datasets. It converts a weak learner (hypotheses) into a better model. It effectively turns a group of several weak learners into a strong learner. These weak learners are used in series. Gradient Boosting has three main components such as loss function, weak learner and additive model. Loss function is the measure of, how far the predicted value is from the original value. Weak learner helps in modelling, whose performance is slightly better compared to random guess. Mostly, decision trees are used as weak learners and the additive model is used for adding the weak learners one at a time. A loss function is being optimized, followed by generation of a weak learner which acts as the predictive model. Every predictor i.e., the weak learner is taught using the predecessor's residual losses as class labels. Then finally the additive model is added to the weak learner and the loss function is minimized. Gradient boosting is an optimal method that quickly over fits the training data. Given a loss function, Gradient Boosting then adopts an additive version, in which it iteratively constructs a chain of estimates in a greedy manner.

The following parameter tuning was performed on the classification model to obtain optimal results. The loss function to be optimized was set to default option as *deviance*; n_estimators that indicate the number of stages of boosting has been set to 100. The Learning rate has been set to 0.1, max_depth that determines the maximum depth of individual estimators has been set to 3 and the random state as 0. The values of these parameters were determined using GridSearchCV, which exhaustively fits in all the possible parameter combinations specified and provides the best parameter combination with the maximum accuracy score. *sklearn.ensemble* Python module with the set parameters has been employed for classifying the LIBS spectral data. Fig. 8 depicts the confusion matrix of polluted SR samples classified by using GBM technique. Using this technique, an overall classification accuracy of 96.43% was obtained, with training and testing accuracy as 100% and 82.14% respectively. The total program execution time for the decision tree classification model is obtained as 581.6s. It is

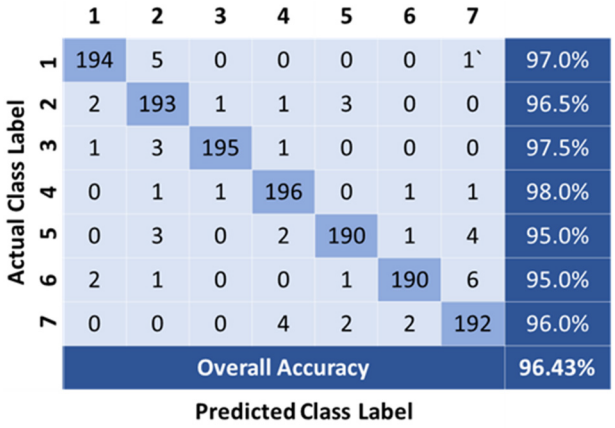

FIGURE 8. Confusion matrix determined by using gradient boosting technique.

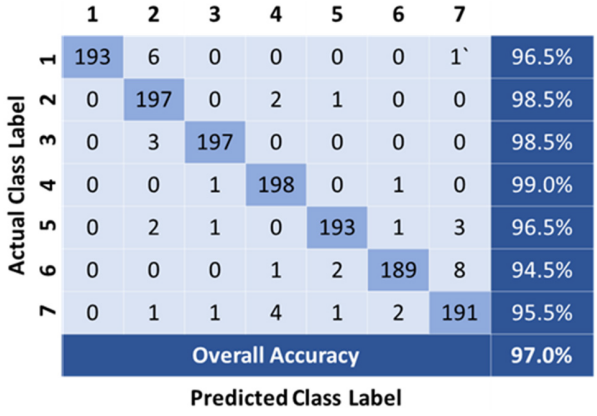

FIGURE 9. Confusion matrix determined by using histogram-based gradient boosting technique.

noticed that the classification accuracy has been improved in case of gradient boosting techniques compared to other methods such as LDA, decision tree and KNN. But, the program execution time has been increased drastically. So, in order to further improve the classification accuracy and to reduce the computation time, advanced gradient boosting techniques have been adopted in the present study and are discussed in the following sections.

5) HISTOGRAM-BASED GRADIENT BOOSTING TECHNIQUE

Gradient boosting has a significant disadvantage because it takes high computation time to train a classifier. This is especially problematic when working with huge datasets. By partitioning (binning) the continuous input values to a few distinct variables, training the trees to the ensemble could be greatly improved. Histogram-based Gradient Boosting, is an ensemble method which adapts histogram-based binning in the traditional gradient boosting framework. The predictors divide the input instances into bins (integer-valued, usually 256 bins), thus greatly reducing the number of splitting points to evaluate and allowing the classifier to create trees using integer-based data sets (histograms) rather than ordered continuous instances. The framework is validated using repetitive stratified k-fold cross-validation, with the mean accuracy score reported over all folds and iterations. The algorithm has an integrated assistance for missing data instances (NaNs); thus, no imputer is required. While training, the tree grower determines if instances with missing data points should be assigned to the left child or the right child depending on the possible gain at each split point. And during prediction, the missing instances are allocated to the left child or right child accordingly.

The following parameter tuning has been performed on the classification model to obtain optimal results. The loss function was set to *categorical_crossentropy*, which is used in case of multiclass classification and max_bins, which determines the maximum number of bins the data instances get divided into, was set to 250.

HistGradientBoostingClassifier from sklearn.ensemble Python module with the set parameters has been employed for classifying the LIBS spectral data. Fig. 9 depicts the confusion matrix of polluted SR samples classified by using histogram-based GBM classifier. Using this technique, an overall classification accuracy of 97.00% was obtained, with training and testing accuracy as 100% and 85.00% respectively. The total program execution time for this classification model is obtained as 174.7s. Using this technique, it is noticed that the classification accuracy has been improved and the program execution time has been reduced when compared to conventional gradient boosting technique.

6) EXTREME GRADIENT BOOSTING (XGBOOST)

Extreme Gradient Boosting is an optimized gradient boosting algorithm. Gradient boosted decision trees are integrated in XGBoost. The decision trees are constructed sequentially in this approach. In XGBoost, the weights are very significant and all of the independent variables are allocated weights, which are subsequently loaded into the decision tree, which predicts outcomes. The weight of factors that the tree estimated incorrectly, is increased and these variables are loaded into the second decision tree. These are then combined to create a more powerful and accurate method for classification. Unlike GBM, XGBoost operates using Newton Raphson method. A cross-validation mechanism is embedded in the XGboost implementation. If the sample data is limited, this mechanism helps the algorithm avoid overfitting. It also includes the *weighted quantile sketch* algorithm, which enables finding the best split coordinates across weighted samples much simpler. The quantiles sketch is a technique for estimating value distribution. In the weighted quantile sketch, these quantiles are weighted so that for each quantile, the sum of the weights within them are almost equal. Furthermore, XGBoost incorporates a specific split finding method called Sparsity-aware Split Finding. This split finding method is capable of handling sparse data.

The following parameter tuning was performed on the classification model to obtain optimal results. max_depth,

which indicates the maximum depth of the tree is set to 5, n_estimators are set to 100, min_child_weight, which allows us to directly control the model complexity is set to default 1 and gamma, which is the amount of loss minimization needed to construct a new split on a tree's leaf node, is set to 1. XGBoost Python module with the set parameters has been employed for classifying the LIBS spectral data. Fig. 10 indicates the confusion matrix of polluted SR samples classified by using XGBoost classifier. With this technique, an overall classification accuracy of 96.92% was obtained, with training and testing accuracy as 100% and 84.64% respectively. The total program execution time for this classification model is obtained as 29.6s. Using this technique, the classification accuracy has been reduced slightly when compared to histogram-based GBM, but the program execution time has been reduced significantly when compared to histogram-based GBM technique.

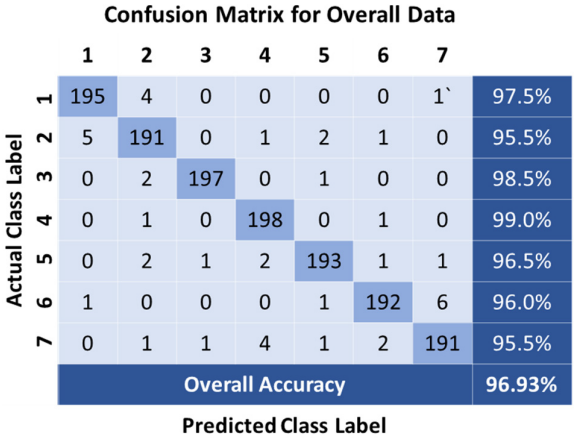

FIGURE 10. Confusion matrix determined by using extreme gradient boosting technique.

7) LIGHT GRADIENT BOOSTING METHOD (LIGHTGBM)

LightGBM is a gradient boosting-based framework which includes efficient data sampling to enable a more accurate and efficient classification model. It uses tree-based algorithms (mainly decision tree) for learning and a histogram-based algorithm (continuous features are grouped into discrete bins and these bins are used to construct feature histograms) for computing the best split. While other boosting algorithms split the tree level-by-level, LightGBM grows the tree leaf-based. It selects the leaf with the lowest delta loss i.e. the leaf is chosen in such a way that it yields the largest decrease in loss. In comparison to the level-wise approach, the leaf-wise algorithm has a reduced loss. Another important feature of this technique is the gradient-based one-sided sampling (GOSS), which is used to sample the data instances by implementing sampling based on the gradients. This technique is used to attain a good balance between reducing the data instances and maintaining the accuracy. The GOSS method deduces that data with a higher gradient contributes more to the information gain. This, along with the Exclusive feature building (EFB), which minimizes the number of features by combining mutually exclusive features are the two novel techniques employed by this method. This is done in order to reduce training complexity resulting in speeding up of training the data and improving prediction. The main difference between LightGBM and other gradient-based methods is that it utilizes GOSS algorithms that divide training samples into smaller subsamples and leaf-wise growth strategy.

The following parameter tuning was performed on the classification model to obtain optimal results. Boosting was set to the default setting *gdbt* (gradient boosting decision trees). num_leaves, which controls the decision tree complexity, was set to 30, num_iterations, the number of iterations were set to 150, max_bin, which decides the maximum number of *feature-buckets*, was set to 60. The values of these parameters were determined using GridSearchCV, which exhaustively fits in all the possible parameter combinations specified and provides the best parameter combination with the maximum accuracy score. LightGBM Python module with the set parameters is employed for categorizing the LIBS spectral data.

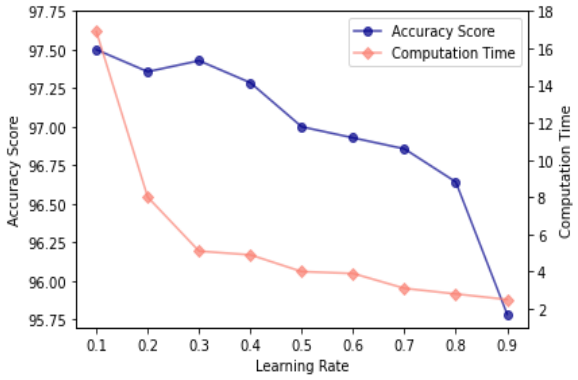

FIGURE 11. Variation in accuracy and computation time with respect to learning rate while using light gradient boosting technique.

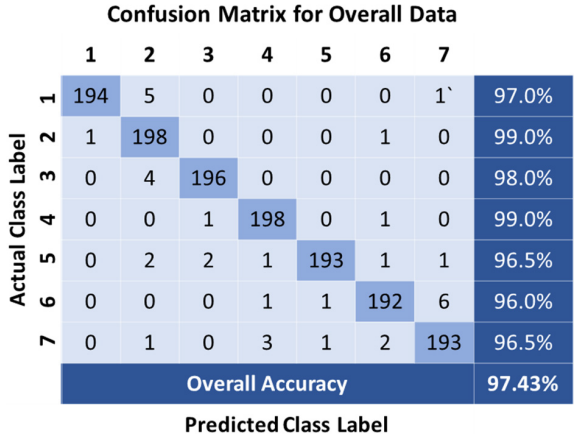

FIGURE 12. Confusion matrix determined by using light gradient boosting technique.

The learning rate has been varied from 0.1 to 0.9 in order to obtain proper classification. The execution time has reduced significantly with the learning rate. But, the classification

accuracy is also noticed to reduce with increase in learning rate. After learning rate of 0.3, classification accuracy is significantly reduced and before 0.3, the execution time is significantly higher (Fig. 11). Therefore, as an optimized value, the learning rate of 0.3 has been selected in the present study for developing a superior classification model. Fig. 12 indicates the confusion matrix of polluted SR samples classified by using LightGBM classifier (with a learning rate of 0.3). An overall classification accuracy of 97.43% was obtained with training and testing accuracy of 100% and 87.14% respectively. The program execution time for the LightGBM classification model is noticed to be 5.1s. The classification model developed using this technique has reflected increased classification accuracy and the significantly reduced program execution time, when compared to the other ML techniques discussed the present study.

C. COMPARATIVE STUDY OF DIFFERENT ML TECHNIQUES

In the present study along with the classification accuracy, the computation time taken each machine learning technique for classifying the test specimens, has been considered for understanding the effectiveness of each ML technique. Table 2 represents the training, testing as well as overall classification accuracies of the polluted SR samples, classified by adopting different machine learning techniques. It is noticed that the KNN method have resulted in an overall accuracy of around 96%, which is significantly higher than LDA and Decision tree methods. The computation time of KNN ids also noticed to be the lowest of all the methods indicating that it is providing faster classification. In order to further improve the classification accuracy, gradient boosting techniques have been employed. The conventional gradient boosting method has shown improved classification accuracy compared to KNN method, but it has resulted in an increased computation time. To achieve higher classification accuracy with lesser computation time, advanced gradient boosting methods such as histogram base gradient boosting, extreme gradient boosting and light gradient boosting techniques have been adopted in the present study. Of all these methods, Light gradient boosting method have reflected higher classification accuracy of 97.43% with a computation time of 5.1 s. Thus, the light gradient boosting technique have achieved improved classification accuracy with reduced computation time. Hence, with the use of LIBS along with machine learning algorithms, pollution deposition on insulators with varied elemental compositions that seem identical to the human eye can be readily classified. Therefore, the predictive ability and the accuracy of LIBS analysis aided by machine learning approaches, might serve as a cost-effective tool to reduce the requirement to perform the laborious experimental techniques on contaminated insulation structures.

TABLE 2. Classification accuracy and computation time taken by various ML methods.

ML Classifier	Training Accuracy (%)	Testing Accuracy (%)	Overall Accuracy (%)	Computation time (s)
Linear Discriminant Analysis (LDA)	100	74.29	94.86	2.2
Decision tree method	100	74.29	94.86	4.0
K-Nearest Neighbors (KNN)	100	80.71	96.14	0.6
Gradient boosting method	100	82.14	96.43	581.6
Histogram-based gradient boosting method	100	85.00	97.00	174.7
Extreme gradient boosting method	100	84.64	96.93	29.6
Light gradient boosting method	100	87.14	97.43	5.1

IV. CONCLUSION

The following are the major conclusions obtained in the current analysis. LIBS analysis is successful in identifying the elemental composition of the various types of pollutants. The presence of copper as well as carbon-based and calcium-based compounds have been identified by the increment in the normalized intensity ratio of Cu I (324.7 nm), C II (657.8 nm) and Ca II (396.8 nm) peaks respectively. LIBS spectral data has been used in conjunction with several ML algorithms such as LDA, decision tree, KNN and various gradient boosting techniques to classify seven different types of contaminated SR samples. When compared to the other ML approaches utilized in this study, classification using the advanced gradient boosting technique i.e. Light gradient boosting technique has reflected better classification accuracy of 97.43% with a computation time of 5.1 s. Hence, the ML assisted LIBS analysis can potentially become a realistic option for a precise classification and speedy analysis of contaminated insulating structures.

REFERENCES

[1] T. Tanaka and T. Imai, *Advanced Nanodielectrics: Fundamentals and Applications*. Singapore: Jenny Stanford Publishing, 2017.

[2] Z. Li, Z. Yang, and B. Du, "Surface charge transport characteristics of ZnO/silicone rubber composites under impulse superimposed on DC voltage," *IEEE Access*, vol. 7, pp. 3008–3017, 2019, doi: 10.1109/ACCESS.2018.2889343.

[3] K. K. Khanum, A. M. Sharma, F. Aldawsari, C. Angammana, and S. H. Jayaram, "Influence of filler-polymer interface on performance of silicone nanocomposites," *IEEE Trans. Ind. Appl.*, vol. 56, no. 1, pp. 686–692, Jan. 2020, doi: 10.1109/TIA.2019.2943445.

[4] T. Han, B. Du, T. Ma, F. Wang, Y. Gao, Z. Lei, and C. Li, "Electrical tree in HTV silicone rubber with temperature gradient under repetitive pulse voltage," *IEEE Access*, vol. 7, pp. 41250–41260, 2019, doi: 10.1109/ACCESS.2019.2907302.

[5] I. J. S. Lopes, S. H. Jayaram, and E. A. Cherney, "A study of partial discharges from water droplets on a silicone rubber insulating surface," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 8, no. 2, pp. 262–268, Apr. 2001, doi: 10.1109/94.919951.

[6] Z. Zhijin, L. Tian, J. Xingliang, L. Chen, Y. Shenghuan, and Z. Yi, "Characterization of silicone rubber degradation under salt-fog environment with AC test voltage," *IEEE Access*, vol. 7, pp. 66714–66724, 2019, doi: 10.1109/ACCESS.2019.2917700.

[7] M. M. Hussain, S. Farokhi, S. G. McMeekin, and M. Farzaneh, "Mechanism of saline deposition and surface flashover on outdoor insulators near coastal areas part II: Impact of various environment stresses," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 24, no. 2, pp. 1068–1076, Apr. 2017, doi: 10.1109/TDEI.2017.006386.

[8] Z. Zhang, X. Qiao, S. Yang, and X. Jiang, "Non-uniform distribution of contamination on composite insulators in HVDC transmission lines," *Appl. Sci.*, vol. 8, no. 10, p. 1962, Oct. 2018, doi: 10.3390/app8101962.
[9] K. Takasu, T. Shindo, and N. Arai, "Natural contamination test of insulators with DC voltage energization at inland areas," *IEEE Trans. Power Del.*, vol. PD-3, no. 4, pp. 1847–1853, Oct. 1988, doi: 10.1109/61.193992.
[10] E. Thalassinakis and C. G. Karagiannopoulos, "Measurements and interpretations concerning leakage currents on polluted high voltage insulators," *Meas. Sci. Technol.*, vol. 14, no. 4, pp. 421–426, Apr. 2003, doi: 10.1088/0957-0233/14/4/303.
[11] A. H. El-Hag, S. H. Jayaram, and E. A. Cherney, "Fundamental and low frequency harmonic components of leakage current as a diagnostic tool to study aging of RTV and HTV silicone rubber in salt-fog," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 10, no. 1, pp. 128–136, Feb. 2003, doi: 10.1109/TDEI.2003.1176575.
[12] P. Charalampidis, M. Albano, H. Griffiths, A. Haddad, and R. T. Waters, "Silicone rubber insulators for polluted environments part 1: Enhanced artificial pollution tests," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 21, no. 2, pp. 740–748, Apr. 2014, doi: 10.1109/TDEI.2013.004015.
[13] N. Wang, X. Wang, P. Chen, Z. Jia, L. Wang, R. Huang, and Q. Lv, "Metal contamination distribution detection in high-voltage transmission line insulators by laser-induced breakdown spectroscopy (LIBS)," *Sensors*, vol. 18, no. 8, p. 2623, Aug. 2018, doi: 10.3390/s18082623.
[14] S. Yang, W. Zhou, J. Yu, H. Li, Z. Rao, J. Lei, and M. Tang, "Influence of aluminum phosphate contaminant on discharge characteristics of suspension glass insulators," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 23, no. 2, pp. 957–964, Apr. 2016, doi: 10.1109/TDEI.2015.005377.
[15] Z. Zhang, D. Zhang, X. Jiang, and X. Liu, "Study on natural contamination performance of typical types of insulators," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 21, no. 4, pp. 1901–1909, Aug. 2014, doi: 10.1109/TDEI.2014.004343.
[16] D. A. Cremers and L. J. Radziemski, *Handbook of Laser-Induced Breakdown Spectroscopy*, 2nd ed. London, U.K.: Wiley, 2013.
[17] W. Xilin, H. Xiao, C. Ping, Z. Chenlong, J. Zhidong, W. Liming, L. Qishen, H. Ronghui, and L. Shungui, "In-situ and quantitative analysis of aged silicone rubber materials with laser-induced breakdown spectroscopy," *High Voltage*, vol. 3, no. 2, pp. 140–146, Jun. 2018, doi: 10.1049/hve.2017.0130.
[18] S. Lu, X. Wang, T. Wang, X. Qin, X. Wang, and Z. Jia, "Analysis of salt mixture contamination on insulators via laser-induced breakdown spectroscopy," *Appl. Sci.*, vol. 10, no. 7, p. 2617, Apr. 2020, doi: 10.3390/app10072617.
[19] J. El Haddad, L. Canioni, and B. Bousquet, "Good practices in LIBS analysis: Review and advices," *Spectrochimica Acta B, At. Spectrosc.*, vol. 101, pp. 171–182, Nov. 2014, doi: 10.1016/j.sab.2014.08.039.
[20] M. A. Gondal, M. H. Shwehdi, and A. A. Khalil, "Applications of LIBS for determination of ionic species (NaCl) in electrical cables for investigation of electrical breakdown," *Appl. Phys. B, Lasers Opt.*, vol. 105, no. 4, pp. 915–922, 2011, doi: 10.1007/s00340-011-4763-1.
[21] X. Wang, S. Lu, T. Wang, X. Qin, X. Wang, and Z. Jia, "Analysis of pollution in high voltage insulators via laser-induced breakdown spectroscopy," *Molecules*, vol. 25, no. 4, p. 822, Feb. 2020, doi: 10.3390/molecules25040822.
[22] V. S. Kumar, N. J. Vasa, and R. Sarathi, "Detecting salt deposition on a wind turbine blade using laser induced breakdown spectroscopy technique," *Appl. Phys. A, Solids Surf.*, vol. 112, no. 1, pp. 149–153, Jul. 2013, doi: 10.1007/s00339-012-7219-5.
[23] H. Rajavelu, N. J. Vasa, and S. Seshadri, "Effect of ambiance on the coal characterization using laser-induced breakdown spectroscopy (LIBS)," *Appl. Phys. A, Solids Surf.*, vol. 126, no. 6, Jun. 2020, doi: 10.1007/s00339-020-03558-7.
[24] W. Li, J. Lu, M. Dong, S. Lu, J. Yu, S. Li, J. Huang, and J. Liu, "Quantitative analysis of calorific value of coal based on spectral preprocessing by laser-induced breakdown spectroscopy (LIBS)," *Energy Fuels*, vol. 32, no. 1, pp. 24–32, Jan. 2018, doi: 10.1021/acs.energyfuels.7b01718.
[25] P. Vinod, M. S. Babu, R. Sarathi, N. J. Vasa, and S. Kornhuber, "Influence of standoff distance and sunlight on detection of pollution deposits on silicone rubber insulators adopting remote LIBS analysis," *IEEE Trans. Ind. Appl.*, vol. 58, no. 3, pp. 3285–3293, May 2022, doi: 10.1109/TIA.2022.3159771.
[26] P. Chen, X. Wang, X. Li, Q. Lyu, N. Wang, and Z. Jia, "A quick classifying method for tracking and erosion resistance of HTV silicone rubber material via laser-induced breakdown spectroscopy," *Sensors*, vol. 19, no. 5, p. 1087, Mar. 2019, doi: 10.3390/s19051087.
[27] S. M. Clegg, E. Sklute, M. D. Dyar, J. E. Barefield, and R. C. Wiens, "Multivariate analysis of remote laser-induced breakdown spectroscopy spectra using partial least squares, principal component analysis, and related techniques," *Spectrochimica Acta B, At. Spectrosc.*, vol. 64, no. 1, pp. 79–88, Jan. 2009, doi: 10.1016/j.sab.2008.10.045.
[28] A. Ren, Q. Li, and H. Xiao, "Influence analysis and prediction of ESDD and NSDD based on random forests," *Energies*, vol. 10, no. 7, p. 878, Jun. 2017, doi: 10.3390/en10070878.
[29] V. C. Costa, F. W. B. Aquino, C. M. Paranhos, and E. R. Pereira-Filho, "Use of laser-induced breakdown spectroscopy for the determination of polycarbonate (PC) and acrylonitrile-butadiene-styrene (ABS) concentrations in PC/ABS plastics from e-waste," *Waste Manage.*, vol. 70, pp. 212–221, Dec. 2017, doi: 10.1016/j.wasman.2017.09.027.
[30] X. Lin, H. Sun, X. Gao, Y. Xu, Z. Wang, and Y. Wang, "Discrimination of lung tumor and boundary tissues based on laser-induced breakdown spectroscopy and machine learning," *Spectrochimica Acta B, At. Spectrosc.*, vol. 180, Jun. 2021, Art. no. 106200, doi: 10.1016/j.sab.2021.106200.
[31] R. Gaudiuso, N. Melikechi, Z. A. Abdel-Salam, M. A. Harith, V. Palleschi, V. Motto-Ros, and B. Busser, "Laser-induced breakdown spectroscopy for human and animal health: A review," *Spectrochimica Acta B, At. Spectrosc.*, vol. 152, pp. 123–148, Feb. 2019, doi: 10.1016/j.sab.2018.11.006.
[32] M. S. Babu, Neelmani, N. J. Vasa, R. Sarathi, and T. Imai, "Use of LIBS technique for identification of type of pollutant and ESDD level on epoxy-alumina nanocomposites using ANN," *Meas. Sci. Technol.*, vol. 32, no. 11, Nov. 2021, Art. no. 115201, doi: 10.1088/1361-6501/ac0d22.
[33] J. E. Sansonetti and W. C. Martin, "Handbook of basic atomic spectroscopic data," *J. Phys. Chem. Reference Data*, vol. 34, no. 4, pp. 1559–2259, Dec. 2005, doi: 10.1063/1.1800011.

K. SANJANA is currently pursuing the B.Tech. degree in data science and engineering with IIT Mandi, India. Her research interests include artificial intelligence, data science, and machine learning and its applications.

MYNENI SUKESH BABU received the Ph.D. degree in high voltage engineering from IIT Madras, Chennai, India, in 2022.

He is currently a Postdoctoral Equivalent Fellow with the High Voltage Laboratory, IIT Madras. His research interests include condition monitoring and design of suitable nanocomposite insulation systems for power apparatus.

RAMANUJAM SARATHI (Senior Member, IEEE) received the Ph.D. degree in high voltage engineering from the Indian Institute of Science, Bangalore, India, in 1994.

He is currently a Professor and the Head of the High Voltage Laboratory, Department of Electrical Engineering, IIT Madras, Chennai, India. His research interests include condition monitoring of power apparatuses and nanomaterials.

NARESH CHILLU (Member, IEEE) received the Ph.D. degree in automation and high voltage engineering from IIT Madras, Chennai, India, in 2021.

He is currently an Assistant Professor with the School of Electronics Engineering (SENSE), VIT AP University. His research interests include automation and preparation of nanocomposite materials for insulation and EMI shielding applications.