MACHINE LEARNING ALGORITHMS FOR CLASSIFICATION OF GAS SENSOR ARRAY DATASET

Rahul.M.G[1], Srujan R Rajanalli [2], Sammed Endoli[3] ,Mahantesh Magi[4],

Dr. N. Ramavenkateswaran[5]

Department of Electronics and Communication Engineering[1,2,3,4], *R V College of Engineering, Bengaluru, India, Assistant Professor,*

Department of Electronics and Communication Engineering[5], *R V College of Engineering, Bengaluru, India*

Abstract: *To measure the accuracy of the data being sensed predictive machine learning models have been used. These models take input in the form of datasets and predict the output based on them. By using large dataset better and efficient predictive models can be designed, because large amount of data can be used to train the model. But having a larger dataset leads to dimensionality problem. This problem is solved using Dimensionality Reduction Principal Component Analysis(PCA) algorithm. PCA helps to reduce the redundant data or correlated data present in the dataset by which dimensionality of the datset is reduced. Classifier algorithms like K Nearest Neighbour(KNN), Logistic Regression(LR), Naive Bayes(NB) and Support Vector Machine(SVM) is used which gives output in the form of confusion matrix. From these confusion matrix the prediction accuracy of models is decided. From the accuracy measurements it is found that SVM model is more accurate(94%) in predicting the output whereas NB model is least accurate(60%).*

Keywords: **Principal Component analysis , K-NN Algorithm ,LR Algorithm , NB Algorithm , SVM Algorithm , Confusion Matrix, K-Fold Cross-Validation.**

1. INTRODUCTION

The demand for extremely accurate gas sensors is growing day by day. To detect various toxic gases a stable device is needed which does not deviate from the accurate response due to environmental factors. Due to its high response time, recovery time and low cost and ease of manufacturing, Metal oxides are being considered an viable sensing material [1]. Metal oxides such as $TiO_2$and $ZnO$are being used to manufacture sensing layers on top of substrate with different metal decorations such as Titanium and Platinum[2]. Methods such as temperature modulation, ultraviolet radiation and advanced fluctuation detection can be used to increase sensing capabilities [3]. Algorithms can be used to improve gas detection accuracy. Some common algorithms implemented are Decision Tree, K Nearest Neighbours, Naive Bayes, Support vector machine, Random forest and Logistic regression. But before the dataset can be given to train the algorithm, a step that can be considered is reducing the dimension. One of the best dimensional reduction techniques is principle component analysis. Reducing the dataset dimension and then using it to model the algorithm may help in increasing the accuracy of classification [4]. Data set of gas sensors response can be obtained from recognised machine learning data repositories. Using these datasets an algorithm can be implemented in real time. In situations where it is difficult to detect gas, like occurance of cross sensitivity and errors due to selectivity, machine learning algorithms and digital processing can be used [5]. This processed data can then be sent through internet of things technology for ease of availability.

2. GAS SENSOR ARRAY DATASET

The data used in this paper is obtained from UCI machine learning repository. This data comprises 13910 observations from 16 chemical sensors that were used in drift compensation simulations in a discrimination challenge involving 6 gases at varied concentration levels. The data was collected in a gas delivery platform facility at the University of California San Diego's ChemoSignals Laboratory in the BioCircuits Institute over a period of 36 months [6]. The resulting collection of dataset includes recordings from six different pure gaseous chemicals, including ammonia, acetaldehyde, acetone, ethylene, ethanol, and toluene, dosed at concentrations ranging from 5 to 1000 ppmv. This dataset

was split into 90% for training and 10% for testing. From the data split of which 10% was allocated for test data, which had around 1390 random observations.

3. ALGORITHMS FOR GAS DATA ANALYSIS

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis is an unsupervised learning approach used in machine learning to reduce dimensionality. With the help of orthogonal transformation, it is a statistical technique that turns observations of correlated features into a set of linearly uncorrelated data. The Principal Components are the newly altered features. It's one of the most widely used programmes for exploratory data analysis and predictive modelling. It's a method for extracting strong patterns from a dataset by lowering variances. It usually attempts to discover the lower-dimensional floor to challenge the excessive dimensional data. PCA works with the aid of using thinking about the variance of every characteristic due to the fact the excessive characteristic indicates the best cut up among the classes, and therefore it reduces the dimensionality.

3.1 K NEAREST NEIGHBOUR ALGORITHM

KNN is an abbreviation of K Nearest Neighbour, which is a supervised learning algorithm and is mainly used to classify data according to the classification method of neighboring data. The algorithm finds the k nearest neighbors of a data point by calculating the distance of all data points. KNN stores all available cases and classifies new cases according to similarity. KNN's K is a parameter that refers to the number next to the most recent included in the majority vote process. KNN is used when the data set is noise less and small. The training data is first loaded. The next step is to choose a value for k. The distance between each row of training data and the test data is determined. Euclidean, Manhattan and Hamming distances can be used to find distance between two points. Euclidean distance is given by,

$$D = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

The values are sorted in ascending order to specify the position of the test points from the top k rows.

3.2 LOGISTIC REGRESSION ALGORITHM

Logistic regression is a classification algorithm estimates the outcome of a classification variable based on an independent variable. It adapts data to logistical functions to predict the probability that an event will occur. The coefficients of the independent variables of the logistic function are optimized by maximizing the likelihood. The maximum likelihood method is a probabilistic framework for solving the problem of density estimation. The decision boundary is optimized so that the cost function is minimal. Gradient descent can be used to minimize the cost function. A sigmod function is used,

$$y_i = \frac{1}{1+e^{-a}}$$

It serves as an activation function in machine learning used to add nonlinearity to machine learning models. Simply put, the value to be passed to the output is determined. A matrix created from the data set is multiplied by a features and is passed to the sigmod function. Then the cost calculation for iteration is done,

$$cosi(w) = (-\frac{1}{m}) \sum_{i=1}^{m} y_i log(y_i) + (1 - y_i) log(1 - y_i)$$

3.3 NAIVE BAYES ALGORITHM

The Nave Bayes method is a supervised learning technique for addressing classification issues that is based on the Bayes theorem. It is mostly utilized in text classification tasks that require a large training dataset. The Nave Bayes Classifier is a simple and effective classification method that aids in the development of fast machine learning models capable of making quick predictions. It's a probabilistic classifier, meaning it makes predictions based on the likelihood of an object. Bayes' law

is a mathematical formula for calculating the probability of a hypothesis given previous information. It is determined by conditional probability. The formula is given by ,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where P(A|B) is Posterior probability ,P(B|A) is Likelihood probability , P(A) is Prior Probability , P(B) is Marginal Probability

3.4 SUPPORT VECTOR MACHINE ALGORITHM

The support vector machine, or SVM, is a widespread supervised learning technique that can be accustomed to solve classification and regression problems. However, it is largely used in machine learning for classification difficulties. The SVM algorithm's purpose is to search out the optimum line or call boundary for categorizing n-dimensional area into categories, so that extra information points will be placed within the correct class in the future without delay. A hyperplane is that name for optimum selection boundary. SVM algorithamic program can be used for face detection, image classification and text categorization.

4. ACCURACY OF ALGORITHM

In this paper , Confusion Matrix and K-fold Cross Validation method is used to determine the accuracy of the algorithms.

4.1 CONFUSION MATRIX

The confusion matrix is a matrix that is used to evaluate the classification models' performance for a given set of test data. Only if the true values for test data are known can it be determined. It's also known as an error matrix since it displays the flaws in the model's performance as a matrix. Accuracy from a confusion matrix is given by,

$$A = \frac{True\ Positive\ Value + True\ Negative\ Value}{Total\ \ Number\ of\ Value}$$

4.2 K-FOLD CROSS VALIDATION METHOD

The K-fold cross-validation method divides the input dataset into K equal-sized groups of samples. Folds are the term for these types of samples. The prediction function uses k-1 folds for each learning set, while the rest of the folds are used for the test set. This strategy is often used in CVs since it is simple to grasp and produces less biassed results than other methods.The following steps are used in this process.The following steps are used in this process

- K groups were created from the input dataset.
- For each group , One group will act as a test dataset and the remaining will act as a training dataset.

5. DESIGN METHODOLOGY

The gas sensor dataset is obtained from UCI machine learning repository and is converted in matrix. Machine learning algorithm KNN, LR, Naive Bayes (NB) and SVM are implemented with python. The matrix dataset is split in training and testing parts. Algorithms are trained using the training data. The design methodology is shown in figure 1. Its accuracy is tested by using confusion matrix and k-fold cross validation with the test data. The algorithm suitable for gas sensing dataset is mentioned.

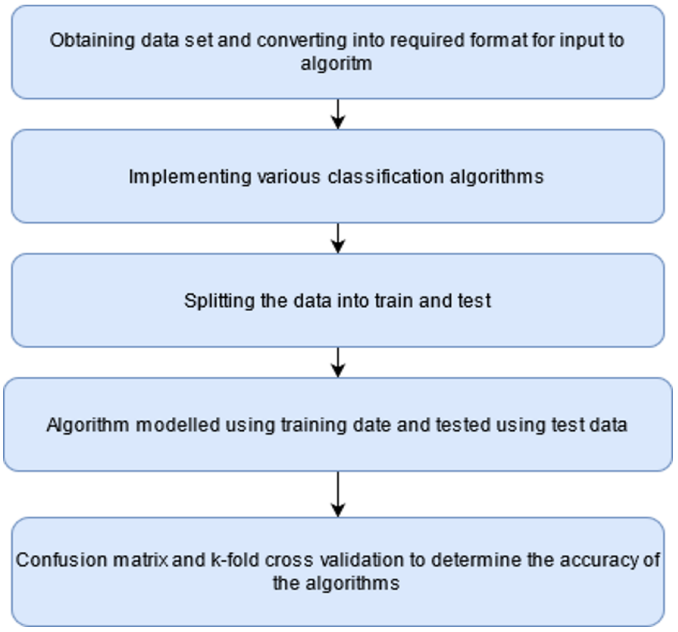

Figure 1. Flow chart of design methodology

6. RESULTS

From the given data set, its dimension was reduced. Figure 2 shows the 3D plot of the principal components after the PCA algorithm is applied. The first component accounts for 70% of the total variance, second and third component accounts for 18% and 5% of the total variance respectively. Using PCA many dimensions can be reduced by eliminating the components which have very little contribution to the dataset.

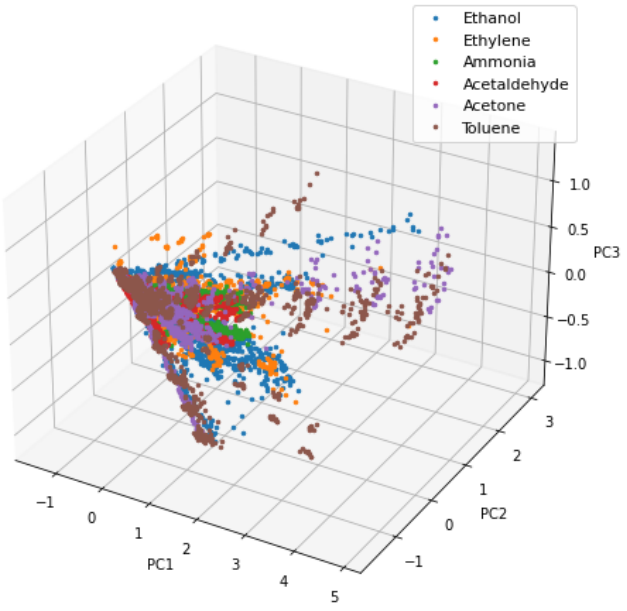

Figure 2. Dimension reduced PCA plot

The accuracy or performance of the classification algorithms can be predicted with the help of a confusion matrix. The diagonal highlighted elements in the confusion matrix are the true predictions by the model, whereas the non diagonal elements are false predictions.

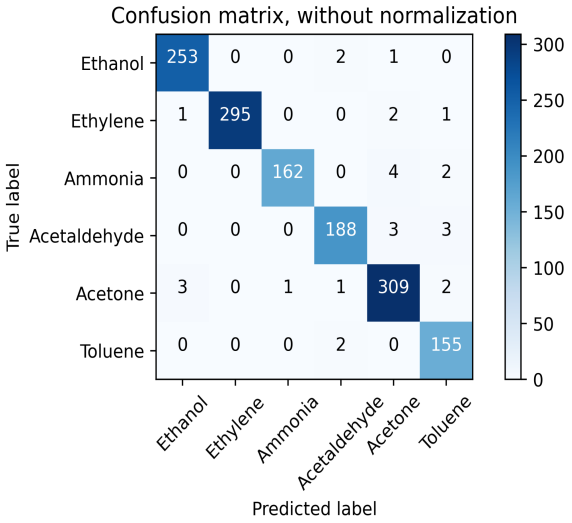

Figure 3. Confusion matrix of KNN algorithm

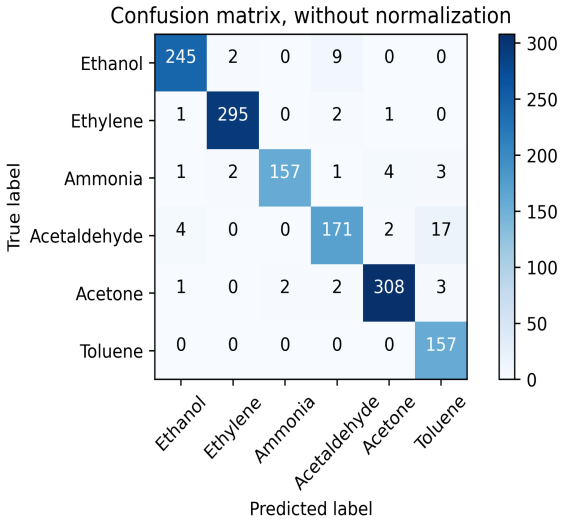

Figure 4. Confusion matrix of Logistic Regression algorithm

Figure 3 shows the confusion matrix for KNN algorithm. The accuracy of the KNN algorithm from the confusion matrix is found to be 97.8%. Figure 4 shows the confusion matrix for logistic regression algorithm. The accuracy of which is found to be 95.8%.

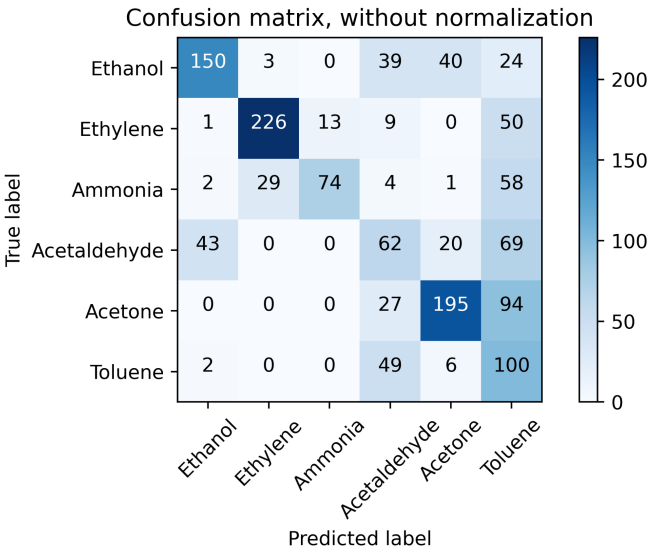

Figure 5. Confusion matrix of Naive Bayes algorithm

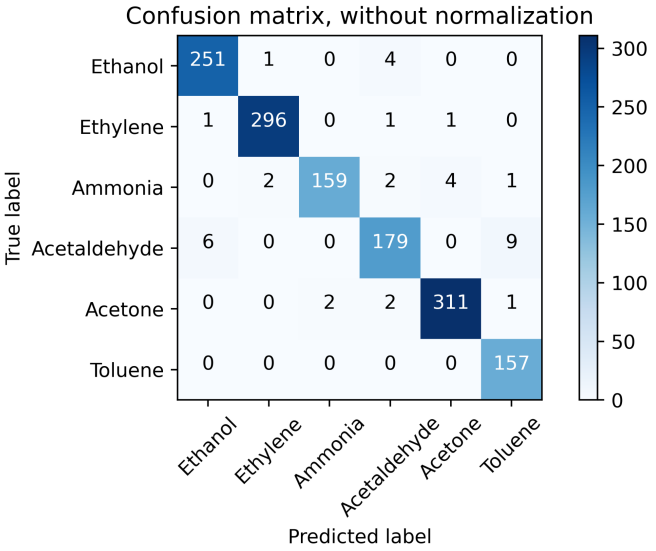

Figure 6: Confusion matrix of Support Vector Machine algorithm

Figure 5 shows the confusion matrix for naive bayes algorithm. The accuracy of which is found to be 58.05%. This algorithm is least accurate in predicting the output compared to other algorithms used in this study. Figure 6 shows the confusion matrix for support vector machine algorithm. The accuracy of this algorithm is found to be 97.33%.

Another approach of predicting the accuracy of algorithms is by k fold cross validation. Using the cross validation points of the algorithms the accuracy is measured. The testing data prediction using k-fold cross validation is shown in figure 7. KNN has a cross validation accuracy of 84.24%. Logistic regression and SVM algorithm have accuracies of 87.3% and 93.6% respectively. Naive Bayes has the least accurate prediction of just 59.7%. From the cross validation approach for validating which algorithm to use, SVM would be preferred along with KNN or logistic regression. SVM is more efficient with higher dimensions of data and when the dimensions are greater than the samples used. SVM also has great memory eciency. The main disadvantage of NB would be that it assumes all predictors (or features) are independent.

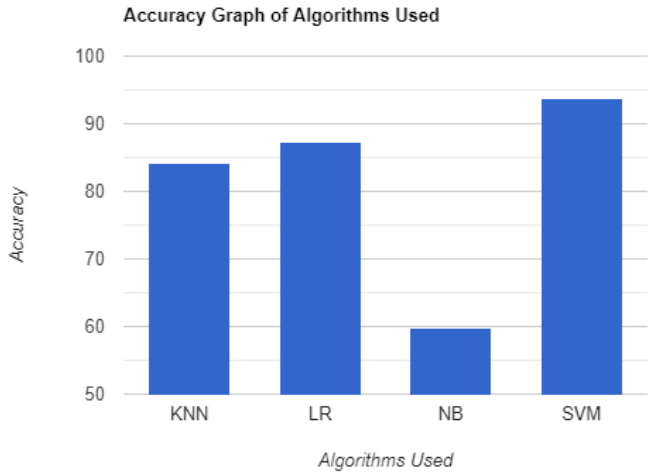

Figure 7: Accuracy graph of different types of algorithms used

7. CONCLUSION

Cross-sensitivity towards interferant gases is unavoidable in metal/metal-oxide-based gas detection. A potential solution to this challenge is to improve gas discrimination by utilising an array of sensors with varied properties. A prerecorded dataset from 16 sensor were used as an input to various classification algorithms. The data set consisted of 13,910 measurement of 6 gases at different concentration whose response was captured by 16 different sensor. This Classification algorithms used were K-Nearest neighbors, Logistic Regression, Naive Bayes and Support vector machine algorithm. A confusion matrix was plotted using the test dataset from which prediction accuracy of all the algorithms were calculated. The results showed that Naive Bayes had the least accuracy with 58.05%. KNN, Logistic regression, and SVM algorithm had an accuracy of 97.8%, 95.8% and 97.33%. Another method for classifying algorithm was used which k-fold cross validation. From this technique SVM had the highest accuracy with 93.6%. KNN and logistic regression had a cross validation accuracy of 84.24% and 87.3%. And the least accurate prediction was by Naive Bayes with just 59.7%. From the results obtained, due to very high prediction accuracy svm can for analyzing different sensor datasets.

8. REFERENCES

[1] *D. Zhang, Z. Yang, S. Yu, Q. Mi, and Q. Pan, "Diversiform metal oxide-based hybrid nanostructures for gas sensing with versatile prospects," Coordination Chemistry Reviews, vol. 413, p. 213 272, 2020, issn: 0010-8545.*

[2] *M. A. H. Khan, B. Thomson, R. Debnath, A. Motayed, and M. V. Rao , "Nanowire based sensor array for detection of cross-sensitive gases using pca and machine learning algorithms," IEEE Sensors Journal, vol. 20, no. 11, pp. 6020–6028, 2020. doi: 10.1109/JSEN.2020.297254*

[3] *J. M. Smulko, M. Trawka, C. G. Granqvist, R. Ionescu, F. Annanouch, E. Llobet, and L. B. Kish, "New approaches for improving selectivity and sensitivity of resistive gas sensors: A review," Sensor Review, vol. 35, no. 4, pp. 340–347, Jan. 2015, issn: 0260-2288. doi: 10.1108/SR- 12-2014- 0747. [Online]. Available: https: //doi.org/10.1108/SR-12-2014-0747.*

[4] A. George, "Anomaly detection based on machine learning dimensionality reduction using pca and classification using svm," International Journal of Computer Applications, vol. 47, pp. 5–8, 2012.

[5] S. Feng, F. Farha, Q. Li, Y. Wan, Y. Xu, T. Zhang, and H. Ning, "Review on smart gas sensing technology," Sensors, vol. 19, no. 17, 2019, issn: 1424-8220. doi: 10.3390/s19173760. [Online]. Available: https://www.mdpi.com/1424- 8220/19/17/3760.

[6] A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta, "Chemical gas sensor drift compensation using classifier ensembles," Sensors and Actuators B: Chemical, vol. 166, pp. 320–329, 2012.

[7] S. Bhattacharya, S. R. K. S, P. K. R. Maddikunta, R. Kaluri, S. Singh, T. R. Gadekallu, M. Alazab, and U. Tariq, "A novel pca-firefly based xgboost classification model for intrusion detection in networks using gpu," Electronics, vol. 9, no. 2, 2020, issn: 2079-9292. doi: 10 . 3390 / electronics9020219.

[8] A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta, "Chemical gas sensor drift compensation using classifier ensembles," Sensors and Actuators B: Chemical, vol. 166, pp. 320–329, 2012.

[9] Y. Xu, X. Zhao, Y. Chen, and Z. Yang, "Research on a mixed gas classification algorithm based on extreme random tree," Applied Sciences, vol. 9, no. 9, 2019, issn: 2076-3417. doi: 10.3390/app9091728.

[10] S. Feng, F. Farha, Q. Li, Y. Wan, Y. Xu, T. Zhang, and H. Ning, "Review on smart gas sensing technology," Sensors, vol. 19, no. 17, 2019, issn: 1424-8220. doi: 10.3390/s19173760.