A Novel Linear Classifier for Class Imbalance Data Arising in Failure-Prone Air Pressure Systems

MUJAHID N. SYED¹, MD. RAFIUL HASSAN², (Member, IEEE), IRFAN AHMAD¹, MOHAMMAD MEHEDI HASSAN³,⁴, (Senior Member, IEEE), AND VICTOR HUGO C. DE ALBUQUERQUE⁵,⁶, (Senior Member, IEEE)

¹Systems Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
²College of Arts and Sciences, University of Maine at Presque Isle, Presque Isle, ME 04769, USA
³College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
⁴Research Chair of Pervasive and Mobile Computing, King Saud University, Riyadh 11543, Saudi Arabia
⁵LAPISCO, Federal Institute of Education, Science and Technology of Ceará, Fortaleza 60040-215, Brazil
⁶ARMTEC Tecnologia em Robótica, Fortaleza 60150000, Brazil

Corresponding author: Mohammad Mehedi Hassan (mmhassan@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Saud University through the Vice Deanship of Scientific Research Chairs: Chair of Pervasive and Mobile Computing.

ABSTRACT An Air Pressure System (APS) is one of the crucial components of an automobile. Its failure leads to financial losses, and it may lead to loss of lives. Thus predicting such failure is a critical problem that requires a rigorous solution. Recently, many researchers have presented machine learning techniques to deal with APS failure detection. One of the major challenges in dealing with APS failure data is the presence of high class imbalance. Conventional classification criteria may not be able to efficiently handle such data. In this paper, a new machine learning method for APS failure detection is proposed. It is designed to specifically deal with the class imbalance. The method uses a linear decision boundary by maximizing Area Under the Curve (maxAUC) criterion. The proposed method was experimentally validated on an industrial dataset of APS failure. The results of the proposed method are thoroughly compared with existing linear as well as non-linear classifiers.

INDEX TERMS APS failure prediction, imbalanced data, area under the curve (AUC), machine learning.

I. INTRODUCTION

As one of the crucial components of an automobile, Air Pressure System (APS) plays a vital role in gauging brakes, shifting gears, adjusting seats, and controlling suspensions. Faulty APS leads to improper functioning of brakes, gears and suspension, which may result in undesired and/or unpleasant incidents. Proper functioning of APS involves efficient (adequate and timely) supply of compressed air to the above systems.

Typically, APS works as follows [1]:

- It cleans and dries incoming compressed air.
- It distributes air to different pneumatic circuits.

Its major components in an automobile are: air drier, circuit protection valves, and control unit. Air drier dehumidifies the incoming air generated at the compressor. The circuit protection valves control the different circuits, like parking brake circuit and service brake circuit, by activating them at different predefined pressures. The control unit consists of temperature and pressure sensors in addition to a circuit board. It decides when to activate the compressor depending on APS pressure level.

APS failures can lead to huge financial losses, and be life-threatening sometimes. Thus, their detection before they actually occur is an important area of research. A critical domain in this research area, known as APS failure detection, is to detect whether an APS failure is the cause of the overall system failure or not. With the advent of Industrial Internet of Things (IIoT) and Industry 4.0, machine-learning based methods for APS failure detection (e.g., [2]) is gaining popularity. One of the major challenges in APS failure detection using machine learning is the presence of high-imbalance class distribution. Generally, the number of samples where APS failure is the key reason for the overall system failure is significantly less than the number of samples where the system failure is due to other reasons. It is widely known that developing a machine-learning based system using imbalance data samples is not straightforward. Based on the literature...
of machine-learning research, imbalanced data can be handled by: modifying the data (like under sampling) to reduce the imbalance, or modifying the machine-learning algorithm (like weighted classification) to handle them. In this paper a new machine-learning method for data classification is proposed that can handle the imbalanced data problem in APS failure detection.

The proposed method is designed to efficiently classify highly imbalanced and overlapping data classes. The presented model uses a linear decision boundary for data classification. The parameters of the decision boundary are obtained by maximizing the Area Under the ROC Curve (AUC). Moreover, the key mathematical model of the proposed method is linear, which leads to a \( O(n) \) worse case complexity. The effectiveness of the method is evaluated on a public dataset of APS failure detection for Scania trucks [3]. The dataset consists of 60,000 samples of failures, where less than 2% of the failures are related to APS. Based on the experiments, it can be seen that our method outperforms previously reported methods on the dataset.

Our contributions can be summarized as:
- Developing a new machine-learning prediction/classification method based on an AUC maximization criterion.
- Incorporating a mechanism to deal with imbalanced data present in APS failure detection.
- The worse case complexity of the proposed method is \( O(n) \).

The rest of the paper is organized as follows: In Section II, related work from the literature is presented. The proposed method and few conventional data classification methods are depicted in Section III. The experiments, results, and discussions are illustrated in Section IV. Finally, the paper is concluded with discussions in Section V.

II. RELATED WORKS

Data analysis and machine learning methods have been frequently applied in diagnosis of transportation systems ([4]–[6]). In this section our focus is to present the work related to APS failure prediction. We also present some important work in the literature related to dealing with imbalanced dataset for classification.

Gondek et al. [7] analyzed APS failure prediction data using multiple existing data classifiers. A number of classifiers including Naive Bayes, Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and random forest were deployed. The authors extracted new features from the APS failure prediction data using feature engineering methods. The original as well as the engineered features were used in the analysis. A feature selection approach based on feature ranking was also used in the analysis. Missing values were replaced by medians in each feature. The results are reported in terms of the cost where predicting a failure wrongly has a cost of 10 units and missing a failure has a cost of 500 units. The authors could achieve an average cost of 0.6 with their proposed technique.

Costa and Nascimento [8] handled APS failure prediction data using weighted data classifiers. In order to deal with class imbalance, class-specific weights were incorporated in Logistic Regression (LogR) and SVM classifiers. The weights assigned to the classes were inversely proportional to the number of samples in a class. In the case of random forest and K-Nearest Neighbors (KNN) classifiers, the authors shifted the threshold for predicting a sample. The threshold was shifted according to the proportion of samples in each class. For example, using KNN, if a minority-class sample was among the 60 nearest neighbors, it was classified as positive. In order to deal with missing data, soft-impute, a form of Expectation Maximization (EM) based imputation, was employed. Results are reported in terms of miss-classification cost, false positive percentage, and false negative percentage.

Peruffo [1] investigated the use of different entropy measures as indicators of node impurity, in order to decide splitting in a decision tree classifier. The author argued that for imbalanced data, alternative definitions of entropy (instead of traditional Shannon’s definition) lead to better measures for splitting. This can lead to improvement in minority-class precision at the expense of majority-class precision. The author tested the entropy measures on the public APS dataset. Accuracy, false negatives, and false positives were used as the performance measures.

Rafsuniani et al. [9] investigated five different imputation techniques (Expectation Maximization, Mean Imputation, Soft Impute, Multiple Imputation by Chained Equation (MICE), and Iterative Singular Value Decomposition (SVD)) in conjunction with five different classifiers: KNN, Naive Bayes, Gradient Boosted Tree, Random Forest, and SVM. Their results indicated that MICE was the most effective imputation technique, and random under-sampling was the most effective technique to deal with imbalanced data. In addition to the accuracy, true negatives, false positives, false negatives, and true positives were used as the performance measures.

Ranasinghe and Parlikad [2] investigated the use of Conditional Generative Adversarial Network (CGAN) in APS failure prediction. They proposed the usage of CGAN for generating artificial samples of the minority class. They used the CGAN by sampling from joint distributions of auxiliary information related to failures and noise. APS failure dataset for Scania trucks were used for the experimentation. The authors generated 2000 extra samples for the minority class to be used during training. True positives, true negatives, false positives, and false negatives were used as the performance measures apart from the misclassification cost.

Akarte and Hemachandra [10] presented the use of gradient boosting trees for APS failure prediction in Scania trucks dataset. The authors employ weighted samples by assigning more weight to the samples of the minority class. The weight was set based on the ratio of the samples of the positive and the negatives classes in the training set. The hyper-parameters were optimized using cross-validation results. Samples with over 70% missing values were removed. Other missing values
in a feature were replaced with feature median. True positives, true negatives, false positives, false negatives, precision, recall, F1-scores were used as measures apart from the misclassification cost.

Jose and Gopakumar [11] presented a modified random forest algorithm for APS fault prediction. Their main contribution was to use bagging with random forest as the training strategy. They also suggested training a separate random forest on the miss-classified data, and use them together with the original random forest classifier. They presented their results on the public APS dataset for Scania trucks. A KNN method with K value of 33 was implemented for imputing the missing values in the dataset. Precision, F-measure, and Matthews Correlation Coefficient (MCC) were used as the performance measures. They reported precision of 0.46 and F-measure of 0.62 on the dataset.

Fatlawi et al. [12] presented a hybrid model for classification problems in the presence of high number of features. They suggested that feature reduction is an important aspect of such problems. Their model is based on K-means clustering and bagging. First, performance for each feature is recorded using a number of metrics. A clustering is then performed on the metric space to partition the space into 2 clusters: one representing a weak cluster and the other a strong cluster. The idea of the former is that the cluster center has values (which represents the specific values of performance metric for different features) lower than the latter’s. The features whose values are members of the weak cluster are regarded as irrelevant features and are removed from further consideration. Finally, bagging-based decision trees are used for training the classification models. The authors report their results on the public APS dataset for Scania trucks. A number of performance measures are used to report the results including precision, recall, F-measure, and AUC. Furthermore, in [13] a feature reduction approach is embedded within the decision tree framework. The node selection is done based on a weighted Gini index value of the feature.

When it comes to dealing with imbalanced datasets, in general, machine learning techniques can be broadly classified into two categories: techniques dealing at the data level (oversampling or under-sampling) and techniques dealing at the classifier level by modifying the algorithms to suit the imbalanced scenario.

Random under-sampling of the majority-class samples, and over-sampling by duplication of the minority-class samples are two of the simplest ways of dealing with the classification of imbalanced datasets. However, they also produce unwanted effects such as over-fitting or information loss by duplicating or deleting examples respectively (cf. [14]). A hybrid technique combines both over-sampling and under-sampling. Synthetic Minority Over-sampling Technique (SMOTE) [15] is another frequently used technique where instances of the minority class are synthetically created between samples of the class and their neighbors. Borderline-SMOTE [14] is a modification of the SMOTE technique, where oversampling of the minority-class is performed only for the samples which are close to the decision boundary. This method considers a minority-class instance to be qualified for oversampling with the SMOTE technique if more than a half of its m nearest neighbors come from the majority class. In [16], under-sampling technique was combined with noise filters in-order to handle noise in the minority class.

Ertekin et al. [17] considered active learning with online SVM to deal with imbalanced data. Online SVM learns incrementally by adding samples one at a time to the training set. The sample to be added to the training set at a given iteration is selected using the active learning strategy. 59 data points are randomly sampled, and the closest sample to the current boundary is added to the training set. An early stopping criterion is used to stop the training. It is based on the idea that if the number of support vectors stabilizes implying that all the possible support vectors have been selected, then the training can stop.

Nguyen et al. [18] presented the idea of oversampling the minority class only at the borderline between the class samples in order to deal with classification for imbalanced data. Their justification was that the samples that lie close to the border are more important for the classification problem. Hence oversampling should be at the borderline instead of using all the minority-class samples. The presented method was found to be effective when the overlap between the classes is low.

Ensemble of under-sampled classifiers is another technique to deal with the imbalanced data (e.g., [19]). As different batches of under-sampled datasets are created, an ensemble seems to perform more robustly as compared to a single classifier. Oh et al. [19] presented an incremental technique based on randomly selecting a balanced subset from the complete data, and then iteratively adding ‘useful’ samples to the training set. The usefulness of a sample is determined by improvement in the information gain of the classifier by adding that sample to the existing subset of training examples. Diez-Pastor [20] created an ensemble of classifiers termed as RB-Boost. The idea was to combine AdaBoost with random sampling. Here, random sampling refers to the idea that proportion of classes in a training set for a AdaBoost instance is selected randomly. Then, SMOTE is used for augmenting data for a class having fewer samples, and random undersampling is used for reducing the number of samples for a class having more data. Through these methods, the goal is to achieve the desired ratio between the class samples.

Shao et al. [21] presented Weighted Lagrangian Twin Support Vector Machine (WLTSVM) for dealing with binary classification of imbalanced data. A graph based under-sampling of the majority-class was presented to deal with imbalanced data. Furthermore, weighted bias was introduced to improve the performance of the class that has fewer samples. Maldonado and López [22] presented a new second-order cone programming formulation for SVM to deal with classification of imbalanced data. The approach is based on
cost sensitive learning, where cost of misclassifying samples of a minority class are higher than that of the majority class samples, and is performed separately for each class. Linear programming SVM formulation was adapted based on second-order cones and the problem was split into two margin variables.

Kang et al. [23] illustrated a Weighted Under-sampling SVM (WU-SVM) method based on space geometry distance. The key idea in WU-SVM, is to generate SubRegions (SRs) by grouping majority samples and assign weights to the majority samples within the SR based to their Euclidean distance to the SVM’s decision plane. This is done in order to retain the data distribution of the original data while under-sampling.

From the literature review, it can be seen that the APS failure detection problem is of importance, and has been considered by many researchers. In addition to that, typical APS data involves imbalanced data. Presence of imbalanced data makes the classification problem a challenging task. In the following section, the proposed method to classify APS failure detection is depicted.

III. RELATED AND PROPOSED METHODOLOGIES

In this section, some of the traditional linear classifiers that can be used for binary classification is presented. Next, the proposed method is given. The method is based on maximizing AUC criterion, and it can handle imbalanced data. Consider classification data \((x_i, d_i)\) for \(i \in \Omega\), containing two classes \(P\) and \(N\), where \(|\Omega|\) is the total observations, \(d_i \in \{-1, 1\}\) is class label, \(x_i \in \mathbb{R}^F\) is the \(i^{th}\) data point in \(F\)-dimensions, \(P \cup N = \Omega\) and \(P \cap N = \Phi\). Without loss of generality, let us say that \(|P| \leq |N|\), \(d_i = 1\) iff \(i \in P\), and \(d_i = -1\) iff \(i \in N\).

A. LINEAR SVM

The basic idea of SVM [24], [25] is to separate two classes (say \(P\) and \(N\)) by a hyperplane defined as:

\[
f(x) = w^T x + b,
\]

such that \(f(x) < 0\) when \(x \in N\), and \(f(x) \geq 0\) when \(x \in P\). Obviously, there could be infinitely many possible choices to select \((w, b)\) in the case of linearly separable classes. Among all these infinite choices, the goal of SVM is to choose \((w, b)\) that minimizes the risk of misclassifying a new unlabeled data point. In other words, the aim is to find a hyperplane that is sufficiently far from both the classes. This can be realized by finding two parallel hyperplanes that separate the classes, such that the following properties are satisfied: the distance (or margin) between the hyperplanes is maximum, and there is no data point in between the hyperplanes. A classifier satisfying the above properties is called a maximum margin classifier. In order to build the maximum margin classifier, without loss of generality, consider the following two parallel hyperplanes:

\[
w^T x + b = 1
\]

Fig. 1 depicts the notion of supporting hyperplanes and the maximum margin. In order to achieve the maximum margin criterion, the following optimization problem is solved:

\[
\text{maximize : } \frac{2}{||w||} \ \
\text{subject to : } y_i(w^T x_i + b) - 1 \geq 0 \quad \forall i \in \Omega.
\]

The distance between the supporting hyperplanes defined in (2) & (3) is given as:

\[
\Delta = \frac{2}{||w||}.
\]

The objective of (5) is replaced by minimizing \(||w||^2/2\), i.e., the above formulation is recast as:

\[
\text{minimize : } ||w||^2/2 \ \
\text{subject to : } y_i(w^T x_i + b) - 1 \geq 0 \quad \forall i \in \Omega.
\]

The above formulations work very well when the data is linearly separable. However, data in most of the practical problems is imbalanced and overlapping. In order to extend the usability of SVMs for overlapping data, additional slack variables are introduced which capture the degree of overlap for some of the data points. This extended classifier is termed as a soft margin classifier, denoted as CSVM, and the changes are incorporated as follows:

\[
\text{minimize : } \frac{1}{2}||w||^2 + c \sum_{i \in \Omega} s_i \ \
\text{subject to : } y_i(w^T x_i + b) - 1 + s_i \geq 0 \quad \forall i \in \Omega,
\]

where \(s_i \geq 0\) is a slack variable, and \(c\) is a parameter that reflects the cost of soft margin. Fig. 2 depicts the notion of a soft margin classifier.
B. LOGISTIC REGRESSION

When data from the two classes overlap, it is sometimes desirable to provide a probabilistic interpretation to the classification results in order to quantify the uncertainty of class labels during prediction. The basic idea of Logistic Regression (LogR) is to assign probabilities to each observation, defined as:

$$h(x_i) = \frac{1}{1 + e^{-w^T x_i - b}} \quad \forall i \in \Omega,$$

(8)

where $w \in \mathbb{R}^F$ and $b \in \mathbb{R}$. The aim of LogR is to choose $(w, b)$ such that $h(x) < 0.5$ when $x \in N$, and $h(x) \geq 0.5$ when $x \in P$. Fig. 3 depicts the probability function.

The optimization model of LogR can be written as:

$$\min \frac{1}{|\Omega|} \sum_{i \in \Omega} \left\| \frac{1}{1 + e^{-w^T x_i - b}} - \frac{d_i + 1}{2} \right\|$$

(9)

The above formulation is recast as:

$$\min \frac{1}{|\Omega|} \sum_{i \in \Omega} \xi \left( \frac{1}{1 + e^{-w^T x_i - b}} \cdot \frac{d_i + 1}{2} \right)$$

(10)

where $\xi()$ is a cost function or a measure of similarity.

C. PROPOSED APPROACH

The basic idea of the proposed approach is to achieve maximum Area Under the Curve (maxAUC) criterion. It can be described through Fig. 4, where $T_{PR}$ denotes the true positive rate, and $F_{PR}$ denotes the false positive rate. The aim of the proposed classifier is to push the $(F_{PR}, T_{PR})$ as close as possible to $(0, 1)$.

Theorem 1: The maximum AUC criterion for binary classification problem can be achieved by maximizing the following function:

$$\max \frac{1}{|N|} \sum_{i \in N} d_i y_i + \frac{1}{|P|} \sum_{i \in P} d_i y_i$$

Proof: AUC can be written as $AUC = 0.5(T_{PR} + T_{NR})$. Assume that the classes are separable. Since $y_i, d_i \in \{-1, 1\}$, a classifier aims to obtain $y_i$ such that $y_i d_i = 1$ for $i \in P$, and $y_i d_i = 1$ for $i \in N$. To sum, an observation $i$ is correctly classified, if and only if, $y_i d_i = 1$ for $i = 1, \ldots, \Omega$.

Now, $T_{PR}$ can be estimated as $\frac{1}{|P|} \sum_{i \in P} d_i y_i$, and $T_{NR}$ can be estimated as $\frac{1}{|N|} \sum_{i \in N} d_i y_i$. Thus, AUC is proportional to:

$$\max \frac{1}{|N|} \sum_{i \in N} d_i y_i + \frac{1}{|P|} \sum_{i \in P} d_i y_i$$

The following LP model can be used to achieve the maxAUC criterion:

$$\max \frac{1}{|N|} \sum_{i \in N} d_i y_i + \frac{1}{|P|} \sum_{i \in P} d_i y_i$$

subject to:

$$d_i (w^T x_i + b) - 1 + s_i \geq 0 \quad \forall i \in \Omega,$$

$$d_i y_i + s_i \leq 1 \quad \forall i \in \Omega,$$

$$y \leq 1,$$

$$y \geq -1,$$

$$s \geq 0,$$

where $y, s \in \mathbb{R}^\Omega$, $w \in \mathbb{R}^F$ and $b \in \mathbb{R}$ are the variables. In addition to that, $D$ is a constant that is estimated from the data points. The objective function in (11) is designed to achieve the maxAUC criterion. Similar parameters and variables are used to compare and contrast the proposed model with SVM and LogR. Constraint (12) and (16) are similar to the soft margin constraint. Furthermore, Constraints (13) to (15) linearly scales the predicted class labels, which is similar to LogR. Fig. 5 depicts a linear scaling function. To sum,
the proposed model combines the characteristics of SVM and LogR, and aims towards improving AUC.

The effectiveness of the proposed model is evaluated in the next section using the standard performance metrics.

IV. EXPERIMENTS, PERFORMANCE METRIC, RESULTS, AND DISCUSSIONS

In this section the performance of the proposed method is experimentally evaluated. At first, a brief description of the performance measures for comparing the results obtained from the proposed and the existing methods is presented. Next, the performance of the proposed method is illustrated by using 2d simulated data. Finally, the APS Scania truck data are classified by using the proposed method, and the results are compared with the existing results.

A. PERFORMANCE METRIC

The following performance metrics are used for evaluating the proposed method and for measuring its effectiveness:

True Positive: True Positive \( (T_P) \) represents the total number of positive samples classified by the classifier that are actually from the positive class. Typically, the minority class is considered as the positive class.

False Positive: False Positive \( (F_P) \) is the total number of samples classified by the classifier as positive while the actual class label of the samples was negative, i.e., from the majority class.

True Negative: True Negative \( (T_N) \) is the total number of samples classified by the classifier as negative that are actually from the negative class.

False Negative: False Negative \( (F_N) \) is the total number of samples classified by the classifier as negative while the actual label of the samples was positive.

Sensitivity: Sensitivity is computed via:

\[
Sensitivity = \frac{T_P}{T_P + F_N} \times 100\% \tag{17}
\]

Specificity: Specificity is computed via:

\[
Specificity = \frac{T_N}{T_N + F_P} \times 100\% \tag{18}
\]

AUC: In addition to the above measures, classifier’s performance can also be evaluated in terms of Receiver Operating Characteristics (ROC). Typically, ROC curve plots sensitivity against specificity for different threshold levels for the classifier. However, for a fixed threshold, ROC can be approximated by a single point on the graph as shown in 4. ROC depicts the performance of a classifier regardless of class distribution. The Area Under the ROC Curve (AUC) summarizes the quality of classification, and is used as a performance measure. AUC is one of the popular performance measures to evaluate a classifier for an imbalanced dataset.

B. EXPERIMENTATION ON SIMULATED IMBALANCED DATA

1) DATA GENERATION

In order to examine the efficacy of the proposed method for classifying binary imbalanced data, a number of synthetic datasets were generated. Specifically, random binary classification data, with various overlapping degrees are generated, where \(|P| = 10\) and \(|N| = 90\). The negative class is Gaussian data with zero mean and unit variance. The positive class is another Gaussian data with unit variance. In Experiment-1, the mean of the positive class is 6 units, which implies a medium level of overlapping. In Experiment-2, the mean of the positive class is 2 units, which implies medium level of overlapping. In Experiment-3, the mean of positive class is 1 unit, which implies high level of overlapping between the samples of the two classes.

2) SOLUTION APPROACH

All the three models, i.e., SVM, LogR, and maxAUC are solved using the standard solution methods. The cSVM model is solved using the libsvm toolbox, the LogR model is solved using glmfit toolbox, and the LP model is solved using the Cplex solver. In addition to that, for cSVM, the hyperparameter \( c \) is cross-validated in powers of 10 from \(-3\) to 3 with an increment of 1. The value of \( c \) that gives the maximum AUC is finally selected.

C. RESULTS AND DISCUSSION ON CLASSIFICATION OF SIMULATED DATA

Fig. 6 shows decision boundaries achieved by all the three methods considered in this study, when no overlapping is present between two classes. As depicted in Fig. 6, for this dataset, the proposed method as well as the other two methods can find a separation boundary between samples of the two classes accurately.

Fig. 7 shows decision boundaries of the three proposed methods in the presence of medium level of overlap between the samples of the two classes. We notice that for the medium overlapping imbalanced dataset, the proposed method can separate the minority class with better sensitivity as compared to LogR and cSVM classifiers, with low trade off in specificity. As shown in Fig. 7, the decision boundary (the green colored line) for the proposed method can classify most of the samples from the proposed and the existing methods is presented.

FIGURE 5. Proposed model scaling.
data points of the minority class (marked as a circle) except one. As mentioned in Section III, for imbalanced data, if a classifier labels all the test data points as the majority class, then the overall classification accuracy would be very high. However, such behavior of the classifier is misleading. From Fig. 7, it can be seen that cSVM and LogR have the misleading behavior. That is, cSVM and LogR push the boundary line to a region such that maximum overall classification accuracy can be achieved. Clearly, the decision boundary of cSVM and LogR is drawn towards the region of the minority class. As a result, most of the data points within the overlapping region are classified as majority class. Table 1 summarizes the training performances of the three methods for the medium overlapping scenario over 100 random trials using the performance metrics considered in this study. For a given measure, and for a given method, average value (standard deviation) of the metric over 100 iterations is presented in Table 1. It is interesting to note that the proposed method was more attentive towards the data points from the minority class, due to the maxAUC criterion. From Table 1, it can be seen that the accuracy of the two conventional methods ($T_P$ for LogR is 7.861; $T_P$ for cSVM is 7.91) in classifying the minority class are worse than our proposed method ($T_P$ using our method is 9.50). However, in terms of the overall accuracy, due to the misleading behavior, it appears that our method has low overall accuracy of 0.9411.

Fig. 8 shows decision boundaries of the three methods in the presence of high degree of overlap between the samples of the two classes. Correspondingly, Table 2 provides the performance summary of the three methods for 100 random trials. As expected, for the highly overlapping dataset, the minority class accuracy of cSVM ($T_P$: 1.05) and LogR ($T_P$: 2.36) are even worse than the case of medium overlap. This is due to the tendencies of the other two classifiers to push the decision boundary towards the minority class samples, with an objective to achieve high overall classification accuracy. The respective AUCs of cSVM and LogR are close to 0.5, which indicates poor performance of these classifiers. Since, the proposed method is based on maxAUC criterion, the accuracy for the minority class classification is very encouraging ($T_P$: 7.961), in the presence of high overlapping and imbalanced data. Furthermore, the high AUC value of 0.78480 also indicates that our method can classify data points from both classes (majority and minority) efficiently. Table 3 shows the average (standard deviation) time per trial in seconds taken by the three methods for the three scenarios.

| Table 1. Classification performances for medium level overlapping classes. |
|-----------------|-----------------|-----------------|
|                | Proposed        | LogR            | cSVM            |
| AUC            | 0.9451 (0.0459) | 0.8863 (0.0686) | 0.8899 (0.0745) |
| Sensitivity    | 0.95 (0.0674)   | 0.786 (0.1326)  | 0.791 (0.1478)  |
| Specificity    | 0.9401 (0.0418) | 0.9866 (0.0084) | 0.9888 (0.01)   |
| Accuracy       | 0.9411 (0.0406) | 0.9665 (0.0184) | 0.969 (0.0181)  |
| $T_P$          | 9.5 (0.6742)    | 7.86 (1.3259)   | 7.91 (1.4777)   |
| $T_N$          | 84.61 (3.7575)  | 88.79 (0.7561)  | 88.99 (0.9045)  |
| $F_P$          | 5.39 (3.7575)   | 1.21 (0.7561)   | 1.01 (0.9045)   |
| $F_N$          | 0.5 (0.6742)    | 2.14 (1.3259)   | 2.09 (1.4777)   |
In this section, the experiments conducted on APS failure prediction is illustrated. First, the demographics of the dataset used for APS failure detection is depicted. Next, a summary of the solution approach is provided. Lastly, results and in-depth comparisons and discussions are presented.

1) APS FAILURE DATASET

APS failure dataset presented in [3] is considered in this experiment. The dataset collected by Scania contains 80,000 data points from heavy trucks operating in five European markets. Each data point is labeled (either positive or negative class), where the positive class represents system failure due to APS failure. 171 different measurements (attributes) are collected from different component of the vehicle, and the attribute names are anonymized by Scania. The aim of this experiment is to predict the APS failure (positive) cases with high accuracy. According to Scania, the cost of incorrectly predicting a non-APS failure as APS failure is $10. Whereas, if APS failure that leads to system failure is undetected, then the incurred cost is $500. The former cost corresponds to APS failure that leads to system failure due to APS failure. In Table 5, the incurred costs for cSVM, LogR and the proposed method is presented. Interestingly, the proposed method can detect failures due to APS with an AUC value of 0.97021 and 0.94339 for training and testing respectively.

2) SOLUTION APPROACH

At first, the missing values per feature is replaced by feature mean, using the KNN method. The value of K based on the literature is taken to be 33. Next, the training data is scaled to zero mean and unit variance. The scaled data is then used for training the models. The solution approach for the APS failure detection data is similar to the simulated data solution approach (Section IV-B2).

3) RESULTS AND DISCUSSION ON APS FAILURE PREDICTION/CLASSIFICATION

Table 4 compares the classification performances of the proposed model with LogR and cSVM. The hyper-parameter (cost C) in cSVM is calibrated by selecting the value that gives the best AUC during the training phase. The proposed method can detect failures due to APS with an AUC value of 0.97021 and 0.94339 for training and testing respectively. Our results clearly outperform the results obtained by LogR and cSVM. As discussed in Section IV-B, for imbalanced data, the overall accuracy is misleading. This misleading behavior is repeated by LogR and cSVM. For the training dataset, cSVM correctly predicts 648 data points out of total 1,000 data points from the positive class. In contrast, the proposed method correctly predicts 967 data points out of 1,000 data points as positive class (i.e., APS failure case). A similar pattern is seen for the test dataset. Evidently, the proposed model can predict the cases of APS failure better than the two conventional classifiers reported in Table 4. For the test data, the number of true positives identified by the proposed method is 345, which is high in comparison with 283 for LogR, and 261 for cSVM. According to Scania, the high true positives achieved not only would minimize the chances of any accidental situations of Scania truck failures, but also would reduce the operational cost. The cost involved in missing any system failure due to APS failure is 50 times the cost of incorrectly attributing system failure to APS failure. In Table 5, the incurred costs for cSVM, LogR and the proposed method is presented. Interestingly, the proposed method can reduce the total cost by a significant amount, when compared to the other two classifiers. As shown in the table, the total amount of savings (if the proposed model is used in predicting APS failure) for training dataset is $144650 when compared with cSVM, and $132439 when compared with LogR. Similarly, the savings for the test dataset is $37290 ($26910) when compared with cSVM(LogR).

Furthermore, training time is a big concern when dealing with big data. Our proposed model is a linear programming model. Hence, the worse case time complexity of the proposed model is $O((\sqrt{n}+d)\log(n+d)\log(d+1))$, where $n$ is the number of training samples, $d$ is the number of class samples is $\approx 1 : 42$, where 375 samples belong to the positive class and 15725 samples belong to the negative class. In addition to that, there are many data points with missing values.

### TABLE 2. Classification performances for high level overlapping classes.

|                | Proposed | LogR   | cSVM   |
|----------------|----------|--------|--------|
| AUC            | 0.7848   | 0.6113 | 0.5515 |
| Sensitivity    | 0.797    | 0.236  | 0.105  |
| Specificity    | 0.7727   | 0.9866 | 0.999  |
| Accuracy       | 0.7751   | 0.9115 | 0.9087 |
| $T_P$          | 7.97     | 2.36   | 1.05   |
| $T_N$          | 69.54    | 88.79  | 89.82  |
| $F_P$          | 20.46    | 1.21   | 0.18   |
| $F_N$          | 2.03     | 7.64   | 8.95   |

### TABLE 3. Time comparisons on simulated data for three classifiers.

|                | Proposed | LogR   | cSVM   |
|----------------|----------|--------|--------|
| No Overlap     | 0.1641   | 0.6481 | 0.0024 |
| Med. Overlap   | 0.1645   | 0.6026 | 0.0083 |
| High Overlap   | 0.1657   | 0.0071 | 0.4023 |

For many pragmatic applications that generates imbalanced data, including the APS failure prediction and classification, it is crucial to correctly classify the samples of the minority class, when compared to correctly classifying the samples of the majority class. The proposed method excels in AUC when classifying synthetically generated imbalanced data. Therefore, we are encouraged to apply the proposed method for classifying APS failure detection data obtained from Scania.

### D. EXPERIMENTS ON APS FAILURE PREDICTION

In this section, the experiments conducted on APS failure detection is illustrated. First, the demographics of the dataset used for APS failure detection is depicted. Next, a summary of the solution approach is provided. Lastly, results and in-depth comparisons and discussions are presented.

For many pragmatic applications that generates imbalanced data, including the APS failure prediction and classification, it is crucial to correctly classify the samples of the minority class, when compared to correctly classifying the samples of the majority class. The proposed method excels in AUC when classifying synthetically generated imbalanced data. Therefore, we are encouraged to apply the proposed method for classifying APS failure detection data obtained from Scania.

### TABLE 2. Classification performances for high level overlapping classes.

|                | Proposed | LogR   | cSVM   |
|----------------|----------|--------|--------|
| AUC            | 0.7848   | 0.6113 | 0.5515 |
| Sensitivity    | 0.797    | 0.236  | 0.105  |
| Specificity    | 0.7727   | 0.9866 | 0.999  |
| Accuracy       | 0.7751   | 0.9115 | 0.9087 |
| $T_P$          | 7.97     | 2.36   | 1.05   |
| $T_N$          | 69.54    | 88.79  | 89.82  |
| $F_P$          | 20.46    | 1.21   | 0.18   |
| $F_N$          | 2.03     | 7.64   | 8.95   |

### TABLE 3. Time comparisons on simulated data for three classifiers.

|                | Proposed | LogR   | cSVM   |
|----------------|----------|--------|--------|
| No Overlap     | 0.1641   | 0.6481 | 0.0024 |
| Med. Overlap   | 0.1645   | 0.6026 | 0.0083 |
| High Overlap   | 0.1657   | 0.0071 | 0.4023 |
features in each sample and $\epsilon$ is the desired accuracy. The time complexity of training cSVM is in the order of $O(nd^2)$ assuming that the number of features per sample are less than the total number of samples in the training set. This is in addition to the time needed to select the optimal value for the hyperparameter $c$. For logistic regression, the time complexity of training is in the order of $O(nd)$ in addition to the number of epochs, i.e., the number of times the algorithm iterates through the complete training set. Hence, the proposed model is not only better in terms of classification performance (as per the AUC measurement) but also in terms of the training time complexity.

As mentioned in Section II, there exist few studies that attempt to predict APS failure using the same dataset. Tables 6 & 7 summarizes the classification performances reported in those studies. It should be noted that, most of the studies use nonlinear classifier to predict the APS failure. In contrast, in this study we propose and develop a new linear classifier built on the maxAUC criterion. Thereby, the performance comparison of the proposed method with the methods reported in the existing studies is biased towards existing studies. Nevertheless, the purpose of the comparison is to benchmark the proposed approach with the existing studies. Tables 6 - 7 provides the performances along with the respective classifiers for APS failure training and testing dataset, respectively. It is encouraging to see that, the proposed model (AUC: 0.94339) is based on linear decision boundary. However, it outperforms Naive Bayes (AUC: 0.9307), KNN (AUC: 0.6364), SVM (AUC: 0.8267), Random Forest (AUC: 0.8075) and Gradient Boosted Tree (AUC: 0.8271) in terms of AUC (results reported in Rafsunjani et al. (2019) [9]). However, as expected, after preprocessing the dataset either by oversampling (e.g., using SMOTE) or under sampling, the performances of non-linear classifiers improve a lot (results reported in Eleonora Peruffo (2016) [1] and Rafsunjani et al. (2019) [9]).

Table 7 also shows that, the proposed method outperforms the AUC of rpart gini (AUC: 0.8418), rpart information (AUC: 0.7962), and J48 (AUC: 0.9035) as reported in Eleonora Peruffo (2016) [1]. It should be noted that, the dataset is first transformed to a balanced one through deploying a SMOTE approach. Thereby, the performances of the classifiers reported in that study are enhanced and Table 7’s row 3 reports the best performances of the respective classifiers. In contrast, the proposed method utilizes linear decision boundary, and maximizes AUC. Hence, we conjecture that no data preprocessing is required to enhance the classification performance. Nonetheless, a nonlinear decision boundary with the maxAUC criterion would certainly provide an enhanced classification performance. As a future work, we plan to develop one such classifier and study its performance on APS failure data.

Rafsunjani et al. (2019) [9] report an enhancement on the performance of a few nonlinear classifiers with an...
under-sampling approach. Table 7’s row 2 shows the best performances of the classifiers in [9]. We see that our proposed model can outperform the performances of Naive Bayes (AUC: 0.9306), SVM (AUC: 0.8411) and Gradient Boosted Tree (AUC: 0.8075) in terms of AUC. All the classifiers reported in Table 7 are nonlinear classifiers except the proposed classifier. AUC values (highlighted in bold fonts) are higher than that of the proposed model. To sum, the proposed linear method can outperform 11 out of 18 different nonlinear models reported in the literature. Although, the performance comparison between linear and nonlinear classifiers on APS data is biased, the proposed linear method empirically proves its capability of handling the complex imbalanced nature of APS failure data. Furthermore, the proposed method also outperforms many nonlinear classifiers, where the data are tailored into a balanced data set for classification. For example, Rafsunjani et al. (2019) [9] report that the AUC of SVM is 0.8411 (after transforming the dataset to a balanced data by using an under-sampling approach). In contrast, the AUC of the proposed method is 0.94339. Generally under-sampling or oversampling approaches have following issues: i) Generating new data samples might introduce impurity to the dataset and ii) valuable information might be lost due to removing samples from majority class. Thus, the sampling approaches to transform imbalanced data into a balanced data might not be always efficient in improving the classification performance. The proposed model aims at maximizing AUC, and thereby it can suitably handle an imbalanced dataset. Thus, it does not require any sampling based transformation of the dataset.

V. CONCLUSION

APS failure detection is a crucial issue in automobile/vehicle systems. The data collected from such systems is typically imbalanced. Thus, the conventional data classification criteria like maximum margin or maximum likelihood may not efficiently handle the imbalanced data. Generally, sampling approaches (under or over sampling) are used to transform the data into a balanced data. The proposed method is designed based on the maximum AUC (maxAUC) criterion. The proposed method by its very nature can handle imbalanced data. Using synthetic data-based experiments, the performance of the proposed method on imbalanced data with varying degree of class overlap is depicted. The proposed method is implemented on the APS failure detection benchmark data collected by Scania. From the performance comparison of the proposed method with the existing methods, it can be stated that the proposed method is a potential candidate for classifying imbalanced data set arising in APS failure detection. In future, we will use deep learning approaches.
on the APS failure detection scenario and evaluate their performances.

REFERENCES

[1] E. Peruffo, “Improving predictive maintenance classifiers of industrial sensors’ data using entropy. A case study,” M.Sc. Research Project, School of Computer Science, National College Ireland, Dublin, Ireland, 2018.

[2] G. D. Ranasinghe and A. K. Parlikad, “Generating real-valued failure data for prognostics under the conditions of limited data availability,” in Proc. IEEE Int. Conf. Prognostics Health Manage. (ICPHM), Jun. 2019, pp. 1–8.

[3] J. Biteus and T. Lindgren, “Planning flexible maintenance for heavy trucks using machine learning models, constraint programming, and route optimization,” SAE Int. J. Mater. Manuf., vol. 10, no. 3, pp. 306–315, Mar. 2017.

[4] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, “Bilevel feature extraction-based text mining for fault diagnosis of railway systems,” IEEE Trans. Intell. Transp. Syst., vol. 18, no. 1, pp. 49–58, Jan. 2017.

[5] D. Yuan, Z. Lu, J. Zhang, X. Li, B. Jia, and T. Ma, “Integrative design of an emergency resource predicting-scheduling-repairing method for rail track faults,” IEEE Access, vol. 7, pp. 155686–155700, 2019.

[6] M. Veres and M. Moussa, “Deep learning for intelligent transportation systems: A survey of emerging trends,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 8, pp. 3152–3168, Aug. 2020.

[7] C. Gondek, D. Hafner, and O. R. Sampson, “Prediction of failures in the air pressure system of scania trucks using a random forest and feature engineering,” in Proc. Int. Symp. Intell. Data Anal. Cham, Switzerland: Springer, 2016, pp. 398–402.

[8] C. F. Costa and M. A. Nascimento, “Ida 2016 industrial challenge: Using machine learning for predicting failures,” in Proc. Int. Symp. Intell. Data Analysis. Cham, Switzerland: Springer, 2016, pp. 381–386.

[9] S. Rafsunjani, R. S. Safa, A. A. Imran, S. Rahim, and D. Nandi, “An empirical comparison of missing value imputation techniques on APS failure prediction,” Int. J. Inf. Technol. Comput. Sci., vol. 11, no. 2, pp. 21–29, Feb. 2019.

[10] M. M. Akarte and N. Hemachandra, “Predictive maintenance of air pressure system using boosting trees: A machine learning approach,” in Proc. 51st Annua. Conf. ORSI Int. Conf. Kolkata, India: ORSI, 2018.

[11] C. Jose and G. Gopakumar, “An improved random forest algorithm for classification in an imbalanced dataset,” in Proc. URSI Asia-Pacific Radio Sci. Conf. (AP-RASC), Mar. 2019, pp. 1–4.

[12] H. K. Patlawi, A. Alharan, and N. S. Ali, “An efficient hybrid model for reliable classification of high dimensional data using K-means clustering and bagging ensemble classifier,” J. Theor. Appl. Inf. Technol., vol. 96, no. 24, pp. 8379–8398, 2018.

[13] H. Liu, M. Zhou, and Q. Liu, “An embedded feature selection method for imbalanced data classification,” IEEE/CAA J. Automatica Sinica, vol. 6, no. 3, pp. 703–715, May 2019.

[14] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A new oversampling method in imbalanced data sets learning,” in Proc. Int. Conf. Intell. Comput. Inform. Conf. Berlin, Germany: Springer, 2005, pp. 878–887.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[16] Q. Kang, X. Chen, S. Li, and M. Zhou, “A noise-filtered under-sampling scheme for imbalanced classification,” IEEE Trans. Cybern., vol. 47, no. 12, pp. 4263–4274, Dec. 2017.

[17] S. Ertekin, J. Huang, L. Bottou, and L. Giles, “Learning on the border: Active learning in imbalanced data classification,” in Proc. 16th ACM Conf. Conf. Inf. Knowl. Manage. (CIKM), 2007, pp. 127–136.

[18] H. M. Nguyen, E. W. Cooper, and K. Kamei, “Borderline over-sampling for imbalanced data classification,” in Proc. 5th Int. Workshop Comput. Intell. Appl., no. 1. Hiroshima, Japan: IEEE SMC Hiroshima Chapter, 2009, pp. 24–29.

[19] S. Oh, M. Su Lee, and B.-T. Zhang, “Ensemble learning with active example selection for imbalanced biomedical data classification,” IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 8, no. 2, pp. 316–325, Mar. 2011.

[20] J. F. Diez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. I. Kuncheva, “Random balance: Ensembles of variable priors classifiers for imbalanced data,” Knowl.-Based Syst., vol. 85, pp. 96–111, Sep. 2015.

[21] Y.-H. Shao, W.-J. Chen, J.-J. Zhang, Z. Wang, and N.-Y. Deng, “An efficient weighted lagrangian twin support vector machine for imbalanced data classification,” Pattern Recognit., vol. 47, no. 9, pp. 3158–3167, Sep. 2014.

[22] S. Maldonado and J. López, “Imbalanced data classification using second-order cone programming support vector machines,” Pattern Recognit., vol. 47, no. 5, pp. 2070–2079, May 2014.

[23] Q. Kang, L. Shi, M. Zhou, X. Wang, Q. Wu, and Z. Wei, “A distance-based weighted undersampling scheme for support vector machines and its application to imbalanced classification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 9, pp. 4152–4165, Sep. 2018.

[24] P. Zhang, S. Shu, and M. Zhou, “An online fault detection model and strategies based on SVM-grid in clouds,” IEEE/CAA J. Automatica Sinica, vol. 5, no. 2, pp. 445–456, Mar. 2018.

[25] S. Teng, N. Wu, H. Zhu, L. Teng, and W. Zhang, “SVM-DT-based adaptive and collaborative intrusion detection,” IEEE/CAA J. Automatica Sinica, vol. 5, no. 1, pp. 108–118, Jan. 2018.

MUJAHD N. SYED received the B.E. degree in mechanical engineering from Osmania University, India, in 2005, the M.S. degree in systems engineering from King Fahad University of Petroleum and Minerals (KFUPM), Saudi Arabia, in 2007, and the Ph.D. degree in industrial and systems engineering (operations research) from the University of Florida, USA, in 2013. He is currently an Assistant Professor with the Systems Engineering Department, KFUPM. His research interests include operations research and data analytics.

MD. RAFLUH HASSAN (Member, IEEE) received the Ph.D. degree in computer science and software engineering from the University of Melbourne, Australia, in 2007. He worked as an Associate Professor with the Department of Computer Science and Information Science, King Fahd University of Petroleum and Minerals and Minerals for more than ten years. He worked with the Department of Computer Science and Software Engineering as a Research Fellow (Lecturer-B), from 2008 to 2010. He is currently working as an Associate Professor of computer science with the University of Maine at Presque Isle (UMPI), USA. His research interests include artificial intelligence, machine learning, and computational intelligence with a focus on developing new data mining and machine learning techniques for big data analysis, cybersecurity, and the IoT. He is the author of around 50 papers published in recognized international journals and conference proceedings. He is a member of the Australian Society of Operations Research (ASOR) and IEEE Computer Society. He is involved in several program committees of international conferences. He also serves as a reviewer of a number of renowned journals, such as BMC Breast Cancer, IEEE TRANSACTIONS ON FUZZY SYSTEMS, Neurocomputing, Applied Soft Computing, Knowledge and Information Systems, Current Bioinformatics, Information Science, Digital Signal Processing, and IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS AND COMPUTER COMMUNICATIONS.

IRFAN AHMAD received the Ph.D. degree in computer science from TU Dortmund, Germany, in 2017. He is currently an Assistant Professor with the Department of Information and Computer Science, KFUPM, Saudi Arabia. He has published several articles in high-quality journals and international conferences. He has also coauthored two book chapters and holds three U.S. patents. His research interests include pattern recognition and machine learning. He regularly reviews articles for well-known journals and conferences in his area of research in addition to being a program committee member for some of the reputed international conferences.
MOHAMMAD MEHEDI HASSAN (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in February 2011. He is currently a Professor with the Information Systems Department, College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia. He has authored or coauthored around more than 260 publications, including refereed IEEE/ACM/ Springer/Elsevier journals, conference papers, books, and book chapters. His research interests include cloud computing, edge computing, the Internet of Things, body sensor networks, big data, deep learning, mobile cloud, smart computing, wireless sensor networks, 5G networks, and social networks. He has served as the Chair and a Technical Program Committee Member in numerous reputed international conferences/workshops, such as IEEE CCNC, ACM BodyNets, and IEEE HPCC. He was a recipient of a number of awards, including the Best Conference Paper Award from IEEE International Conference on Sustainable Technologies for Industry 4.0 (STI) 2020, the Best Journal Paper Award from the IEEE SYSTEMS JOURNAL, in 2018, the Best Paper Award from CloudComp Conference, in 2014, and the Excellence in Research Award from King Saud University (two times in 2015 and 2016). He is one of the top 2% Scientists of the world in Networking and Telecommunication field. He is one of the top computer scientists in Saudi Arabia as well. Recently, his four publications have been recognized as the ESI Highly Cited Papers.

VICTOR HUGO C. DE ALBUQUERQUE (Senior Member, IEEE) graduated in mechatronics engineering from the Federal Center of Technological Education of Ceará (CEFETCE), in 2006. He received the M.Sc. degree in teleinformatics engineering from the Federal University of Ceará (UFC), in 2007, and the Ph.D. degree in mechanical engineering from the Federal University of Paraíba (UFPB), in 2010. He is currently a Professor and a Senior Researcher with the LAPISCO/IFCE and ARMTEC Tecnologia em Robótica, Brazil. He is a specialist, mainly, in the Internet of Things (IoT), machine/deep learning, pattern recognition, and robotics.