2021 24th International Conference on Computer and Information Technology (ICCIT), 18-20 December, 2021.

2021 24th International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-9435-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCIT54785.2021.9689842

IRFD: A Feature Engineering based Ensemble Classification for Detecting Electricity Fraud in Traditional Meters

Md. Zesun Ahmed Mia
Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology,
University of Liberal Arts Bangladesh
Dhaka, Bangladesh
zesun123@gmail.com

Md. Moinul Islam
Computer Science and Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
moinulislam7002@gmail.com

Monjurul Haque
Electronics and Telecommunication Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
mhshakib07@gmail.com

Saiful Islam
Computer Science and Engineering
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
islam.saiful03@outlook.com

S.M.A Mohaiminur Rahman
Mechanical Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh
shovon2021@hotmail.com

***Abstract*—Nations have suffered significant economic losses as a result of non-technical electric losses resulting from power fraud. It is a criminal act of stealing electricity by applying various mechanisms that incorporate unauthorized tapping to the power line, bypassing the smart meter, etc. Electricity theft is a significant concern for not only developing countries but also developed countries as well. However, for most developing countries, the implications are catastrophic, given that their usage is always less than their demands. Electricity theft must be detected precisely and quickly in order to be mitigated. In our study, we have proposed a method of predictive ensemble machine learning techniques (IRFD) with a novel combination of feature distinction methods to detect electricity theft. In our proposed model, we have combined feature selection technique, Recursive Feature Elimination with Stratified 10-Fold cross-validation (RFECV) and Isolation Forest (IF), to identify and remove outliers along with several machine learning classifiers to forecast the theft of electricity. This study additionally enhances the management of highly imbalanced fraudulent data with Borderline-SMOTE with SVM (SVMSMOTE) and feature scaling with StandardScaler. Following the study, the Random Forest classifier observed a higher degree of accuracy (97.06%) with higher precision, recall, and F1-Score. To evaluate the efficacy of our proposed model, comparative analysis of the classification metrics is also assessed with several machine learning classifiers like Logistic Regression, Gradient Boosting, XGBoost, AdaBoost, KNN, ANN, along with Random Forest before and after fitting our proposed feature engineering techniques.**

***Index Terms*—Electricity Fraud, Machine learning, Anomaly detection, Isolation Forest, Recursive Feature Elimination with Stratified 10-Fold cross-validation (RFECV), Random Forest.**

I. Introduction

The most essential gift that science has given mankind is electricity. Electricity is essential in our everyday lives. It is critical not just in our homes but also in businesses, hospitals, and communication networks, among other places. As a result, during the last several decades, energy consumption has been quickly rising, necessitating more electricity to satisfy humanity's needs. During the production, transmission, and distribution of electricity, there are losses [1]. These losses can be categorized into two parts. One is the technical loss caused by electricity energy dissipated in heat, such as core loss and iron loss, which occurs between the source and the consumer's distribution point. Another type of loss is non-technical loss, including electricity theft at the distribution level, metering problems, and billing errors. electricity power theft may take several forms, including tapping a line or circumventing the energy meter. According to research, individual residences account for 80% of theft globally, while commercial and

978-1-6654-9435-9/21/$31.00 ©2021 IEEE

industrial properties account for 20% [2]. Electricity theft is a massive problem in the current electricity grid, resulting in significant financial loss. It costs $6 billion a year in the United States, $173 million in the United Kingdom. A loss of roughly $100 million each year occurs in Canada. The losses have far more significant effects on developing countries [3]. To mitigate this problem, electricity theft needs to be detected. Many studies have been conducted to identify electric theft utilizing various approaches such as machine learning, deep learning, and others.

There have been quite a few researches on electricity theft detection in Smart Grids, mainly using AMI (Advanced Metering Infrastructure) [4]–[7] and various ML (machine learning) techniques [8]–[10]. However, beyond the advantage of smart meters, the research works involving electricity theft using ML have yet to be established in the scientific community because the data obtained from concerned authorities are neither clean nor diverse enough to introduce many features while applying ML algorithms. The lack of features and presence of erratic and unmethodical data pose difficulties in applying ML algorithms. Razavi et al. [11] has reported the impact of using various machine learning (ML) models in the case of six different attack types in a smart grid. They developed a feature engineering framework based on a Finite Mixture Model clustering model coupled with an evolutionary Genetic Programming method, and their approach is based on the AMI. Sasirekha et al. [2] used appropriate machine learning methods, such as decision tree and random forest driven SVM, to estimate the anticipated electricity consumption of distribution level customers. However, he only showed a prediction accuracy of RF and DT-based SVM to be 84.4%, 88.52% respectively. The K-nearest neighbors (KNN) classifier was chosen in [1] because of its speed and simplicity; nevertheless, the system was able to identify power theft with only a 91.00% accuracy. The multitask feature extraction fraud detector (MFEFD) is a deep-learning-based model for identifying electricity theft in smart metering infrastructure. Hu et al. [12] claimed that MFEFD's strong feature extraction capacity was due to its deep structure and high nonlinearity. MFEFD was trained in a semisupervised way, allowing it to use both unlabeled and labeled data, overcoming the challenges of a lack of data resources for power distribution system fraud detection.

In this work, we have proposed an ensemble approach (IRFD) to identify non-technical losses such as power theft or fraud using feature distinction and machine learning approaches. Machine learning classifiers with a novel feature engineering combination are used in our approach. The dataset is obtained from the Tunisian Company of Electricity and Gas (STEG) [13]. We have pursued the approach of cleaning and categorizing the data before applying the ML algorithms. To begin our research, the data is preprocessed to include the features, and a significant number of new features are generated. This feature creation allows the machine learning algorithm to learn more quickly and readily from the supplied data to anticipate fraud instances. After that, "Level Encoding" is used to transform the categorical data into numerical data. Furthermore, the "dropna" function is used to impute the missing value from the dataset. "StandardScaler" is then applied to scale the features, and "Isolation Forest (IF)" is utilized to identify and eliminate anomalies in the data in order to enhance the quality. After that, The data balancing is then carried out using "SVMSMOTE," which focuses on detecting new instances of minority class near borderlines using SVM. In addition, to improve the model's accuracy. Recursive Feature Elimination with cross-validation (RFECV) is employed for feature selection to determine the most impactful features responsible for detecting fraudulent cases. In our research, 70% of the data is set aside for training, 20% for testing, and 10% for validation purpose. Following that, we have utilized the "Random Forest (RF)" in our method to identify energy theft and observed a substantial improvement in the evaluation metrics for detecting electricity fraud. Finally, a comparison of several machine learning classifiers such as Artificial Neural network (ANN), Gradient Boosting Classifier (GBC), K-Nearest Neighbor (KNN), Logistic Regression (LR), and Extreme Gradient Boosting (XGBoost) along with Random Forest (RF) are performed to demonstrate the efficacy of these classifiers and choose the best one. Also explored is the effect of utilizing feature distinction techniques like Recursive Feature Elimination with cross-validation (RFECV) and IF in the proposed model for the aforementioned algorithms.

II. Methodology

In this section, we have explained the overall architecture that detects non-technical electricity theft. On a larger scale, our proposed framework mainly incorporates two main stages, data pre-processing and the machine learning pipeline.

A. Data Preprocessing

Data preprocessing entails converting unstructured data into well-formed data sets. Raw data is frequently incomplete and formatted inconsistently. Data preprocessing ensures the dataset is prepared for interpretation and can be parsed by machine learning algorithms.

At first, we ensure that all of the data are of the same type of variables; for that, we have used 'Label Encoding'. It is one of the most standard approaches for dealing with categorical variables. Label encoding is the process of transforming categorical data into machine-readable numeric form. A numeric value has been substituted between 0 and the number of classes minus 1 for the category value like, if a categorical value has four unique classes, then the numeric values will be (0,1,2,3). To remove the missing values, row-wise/column-wise (such as Null/NaN/Nat) from the dataset, pandas' default function "dropna" has been used.

B. Machine Learning Pipeline

This section emphasizes the robust feature engineering methods and machine learning models to classify and predict the losses and reduce the misuse of electricity data more accurately and effectively. 70% of the data are used for

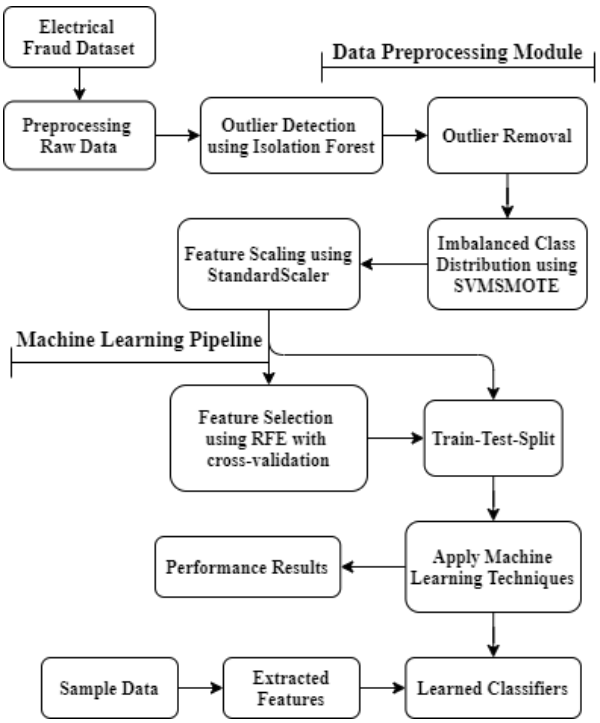

Fig. 1. Proposed Methodology

training, 20% for testing, and 10% for validation purpose. The overall methodology of our proposed model is illustrated in Figure 1

1) Anomaly Detection with Isolation Forest: Isolation forest is one of the finest ways for detecting anomalies in high-dimensional data sets [14]. It is a tree-based method specifically developed to find outliers. It is based on the concept of creating a random forest. Existing data is used to make decisions. The decision tree grows at random with each decision made. At each node, a feature is chosen at random. To split the dataset in half, a random threshold value (between the minimum and maximum) is also chosen. As the process progresses, each observation becomes increasingly isolated from the others. Anomalies are data that differ significantly from the rest of the data. As a result, they are segregated in fewer steps than when observed usually. Thus, to make data easier to analyze and learn from, we have applied the isolation forest method in our research.

2) Imbalanced Class Distribution using SVMSMOTE: Imbalanced class distribution is such a prevailing problem. It is a case where the number of observations in one class is considerably smaller than the number of observations in the other classes, which consequently causes inaccurate and biased ML models. To surmount this, we have adopted SVMSMOTE. It is a variant of Synthetic Minority Oversampling Technique (SMOTE) [15], [16] algorithm based on a Support Vector Machine (SVM) algorithm to identify samples to be used in the formation of new synthetic samples. Its forerunner, SMOTE is a common oversampling strategy that balances class distribution by synthetically generating new minority class instances along with directions from current minority class examples towards their nearest neighbors. SVMSMOTE employs Support Vector Machine to generate new minority class instances by interpolating around borderlines, assisting in establishing class boundaries.

After that, we have applied the StandardScaler as a feature scaling method on our dataset. Being one of the most important steps during data preprocessing, it is a standardization technique to scale the independent features to bring them in the same fixed range. That's why, the mean is set to 0 and the data is scaled to unit variance. Given the data distribution, the mean value will be removed from each value in the dataset, which will then be divided by the whole dataset's standard deviation. Furthermore, in the case of multi-variate data, this is done independently feature-wise for each column of the data.

3) Feature Selection using RFE with Cross Validation: Moreover, in order to distinguish important features that will represent the dataset well enough and be contributory to the ML algorithm, excluding irrelevant, redundant, and noisy features, we have utilized Recursive Feature Elimination with cross-validation (RFECV). It is a wrapper-type feature selection algorithm [17].

RFE uses a supervised learning estimate that has previously been fitted to all features using data. The coefficient associated with each feature is then taken into account. The feature with the smallest absolute coefficient value is deemed the least significant, and so on. The least important coefficient is then removed from the list of features, and the model is rebuilt using the remaining features. It is preferable to remove one feature at a time since the coefficient values of other features change when the model is rebuilt. It rebuilds the model with each iteration, removing the least significant feature(s), and repeating the process until it only has the most significant features, and it turns out to be 95 for our dataset. After that, it rates features based on how long they took to be eliminated. The highest rank is given to the feature that was deleted first, and so on.

III. Implementation & Performance Evaluation

This section defines the performance metrics that are used in this study to evaluate all classification models in terms of non-technical electricity fraudulent data. We have used evaluation metrics such as, *Precision*, *Recall*, *Accuracy*, *F1-Score* to show the experimental results.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{1}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Our proposed method consists of different modules. This model has been developed on a machine having OS Windows

10, 512GB SSD, Core i7-10750H CPU @2.60GHz, 16GB RAM, and Nvidia Geforce GTX 1660Ti GPU. The system has been developed in Python 3.7, in which numpy, scikit-learn, tensorflow, and other packages are used to implement this method.

A. Outlier Detection & Removal

We have initially processed the raw data set for integration into the outlier identification model. as previously stated in Section II. At first, the Label encoder transforms all category information into numerical values. Then, all the features are normalized using StandardScaler. Some characteristics showed much better distribution after the scaling than in the previous one, as seen in Figure 2.

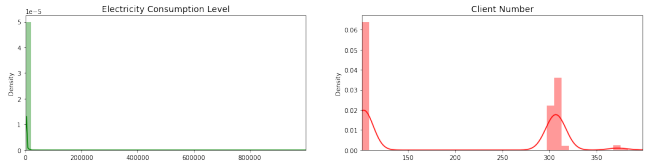

Fig. 2. Data Distribution of Sample Features after Feature Scaling

Data is subsequently adapted to the Isolation Forest model. Isolation forest separates each point in the data, and an anomalous point is separated in a few steps, with specific, nearby points taking much more steps towards separation. If a normal instance exists, it assigns a value of 1 to an instance; otherwise, -1 is assigned if it exists. For roughly 5% of our dataset, we have categorized anomalies based on isolation forest. Once the anomalies have been detected, we eliminated them from our datasets to enhance classification precision. To illustrate the outcome, we have normalized and adjusted the measurements to PCA to decrease the number of dimensions and then plotted them in 3D, highlighting the anomalies in Figure 3.

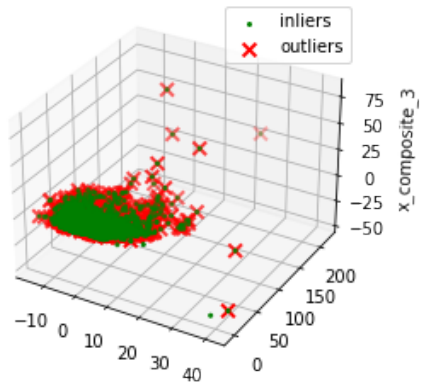

Fig. 3. 3D Scatter Plot of Outlier Detection

B. Data Balancing using SVMSMOTE

After we have detected and deleted outliers using Isolation Forest from our dataset, we have evaluated whether our dataset is balanced or not. However, the dataset is significantly unbalanced, with 6,765 instances, of the 11,340 cases where fraudulent cases are discovered, 5.8 percent of the total cases displayed in figure 4a. We have used Borderline-SMOTE (SVMSMOTE) to handle a randomly selecting minority class instances, replacing them, and adding them to the training data set and SVM's method identifies misclassified cases on the minority class decision boundary along the lines that connect with interpolation each of the vectors supporting minority classes with their nearest neighbors. The number of non-fraudulent and fraudulent instances turns out to be 77428 for both 0 and 1 as shown in Figure 4b.

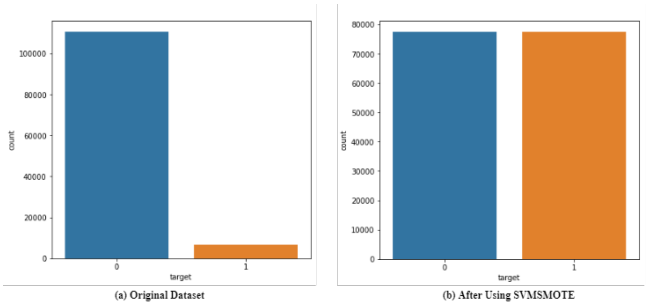

Fig. 4. Data Representation of "Target" Feature (0 represents Fraud, 1 represents Normal)

C. Feature Selection using RFECV

We were challenged to cope with a reasonably broad range of data. Features that are not relevant might increase the run-time of the system while giving disappointing results. We have utilized the Recursive Elimination Feature to choose a subset of the most important features in our data set (RFE). You can pick these characteristics (columns) more or more important to predicting the target variables in a training data set. Fewer capabilities make it easier for machine learn algorithms to work with less time or space complexity. Some algorithms may be misled by irrelevant inputs, leading to worse prediction efficiency. There are 104 characteristics in our dataset. All characteristics do not contribute to the classification equally. After performing Recursive Feature Elimination with Cross-Validation Technique (RFECV), 95 out of 104 features have been identified as the most relevant. We have utilized the approach of the "Stratified 10-Fold cross-validation" technique for cross-validation to verify that in the initial data, training set and test set, the corresponding proportion is the same.

D. Applying Machine Learning Algorithms

In this research, we have applied several machine learning algorithms such as K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), Gradient Boosting Classifier (GBC), XGBoost Classifier (XGB), AdaBoost Classifier (ADA), and Artificial Neural Network (ANN). At first, a comparison of accuracy is made among these algorithms after preprocessing the dataset. After that, we have observed the classification accuracy of electricity fraud detection after

eliminating irrelevant features by RFECV, and the accuracy has improved significantly for the classification algorithms except for Logistic Regression as illustrated in I. The proposed method also demonstrates improved performance in terms of Precision, Recall, and F1-score values. For a better view, a graphical representation of accuracy for all the algorithms applied has been illustrated in Figure 5.

TABLE I
ACCURACY, PRECISION, RECALL, F1-SCORE

	Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
After Pre-processing	KNN	85	96	88	92
	RF	94.03	96	98	97
	LR	83.87	97	86	91
	GBC	92	96	86	91
	XGB	**94.2**	96	98	97
	ADA	88.98	96	92	94
	ANN	90	96	93	95
After Feature Selection (RFECV)	KNN	95.87	98	94	96
	RF	**97.06**	96	98	97
	LR	80.45	78	86	81
	GBC	93.22	91	95	93
	XGB	96.96	96	98	97
	ADA	90.21	89	92	90
	ANN	93	94	92	93

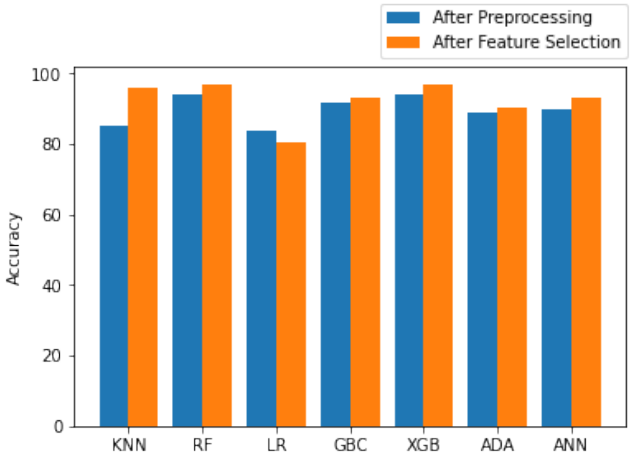

Fig. 5. Bar Chart of Classification Accuracy

1) Result Analysis on the Preprocessed Dataset: In the methodology section, we have mentioned that we have re-sampled the dataset using SVMSMOTE and removed the anomalies on the dataset using the Isolation Forest algorithm. From Table I, we observe that the algorithms KNN, RF, LR, GBC, XGB, ADA and ANN have accuracy of 85%, 94.03%, 83.87%, 92%, 94.2%, 88.98% and 90% respectively for the preprocessed dataset. Among the classifiers, Extreme Gradient Boosting Classifier (XGB) has the highest accuracy of 95% among other algorithms, and Logistic Regression (LR) has the lowest accuracy of 83.87%. In contrast, LR has obtained the best precision rate of 97%. Both RF and XGB have attained the highest recall and F1-Score of 98% and 97%. In the case of ANN, we have implemented this algorithm with 150 epochs and a batch size of 20. We also have evaluated the loss and accuracy after implementing ANN in our model. The loss was high initially but started reducing after a certain period, and the accuracy was almost constant throughout the epochs as illustrated in Figure 6a.

Moreover, to evaluate the performance of our proposed model, the Receiver Operating Characteristic (ROC) curve of all the classifiers for the preprocessed dataset is shown in Figure 7, where XGB and RF have better area under the ROC curve of 0.85 and 0.849 respectively among other algorithms.

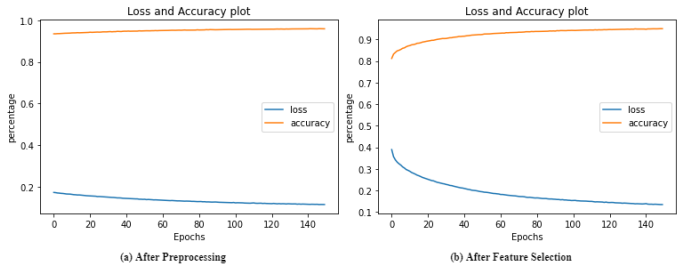

Fig. 6. Loss and Accuracy Plot for ANN

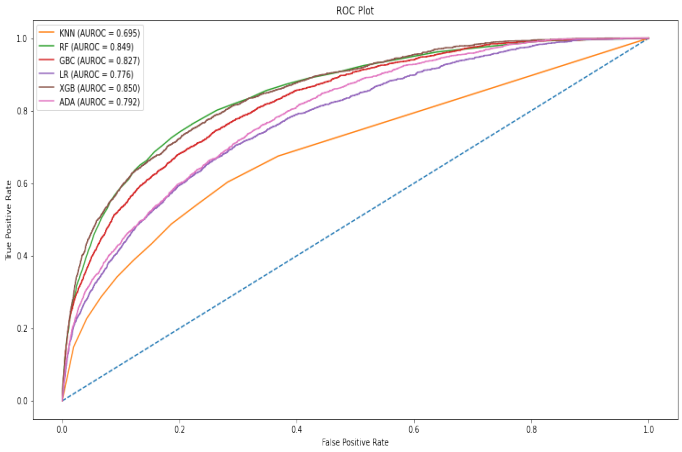

Fig. 7. ROC Curve for Machine Learning Algorithms (After Preprocessing)

2) Result Analysis after Feature Selection: Using the Recursive Feature Elimination with cross-validation (RFECV) technique, we have removed the irrelevant features to improve the classification accuracy of the machine learning algorithms. We observe that the electricity fraud detection model's classification accuracy has improved substantially after applying RFECV accuracy of 95.87%, 97.06%, 93.22%, 96.96% and 90.21% respectively for KNN, RF, GBC, XGB, ADA except for Logistic Regression (80.45%). The proposed method shows improved performance in terms of Accuracy Precision, Recall, and F1-score values. Table I shows that Random Forest (RF) classifier attains the highest accuracy, precision, recall, and f1-score of 97.06%, 96%, 98%, and 97% respectively and outperforms other classification algorithms applied, whereas the accuracy rate (96.96%) of XGB also shows a promising

result. After analyzing the precision rate, it can be observed that KNN has the highest precision rate of 98% than other classifiers. Apart from these, ANN gives better performance after feature selection in terms of accuracy with 93% than that of the preprocessed dataset as shown in Figure 6b.

With the area under the ROC curve, we can also evaluate the changes before and after applying the feature selection module, RFECV as shown in Figure 8. Here, RF and XGB have attained better AUROC scores of 0.993 and 0.992 respectively than other classifiers.

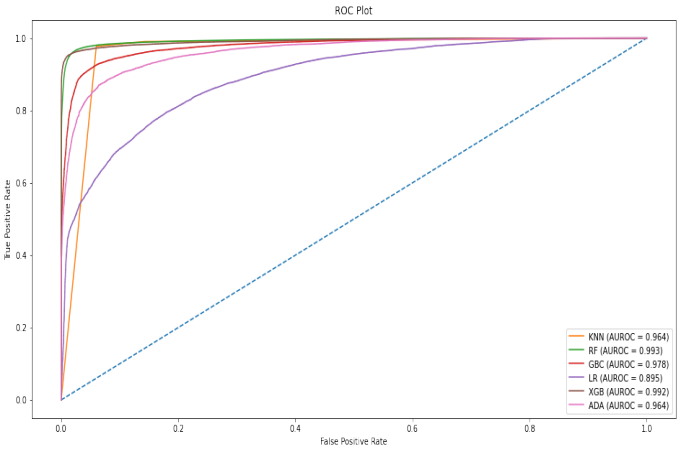

Fig. 8. ROC Curve for Machine Learning Algorithms (After Feature Selection)

IV. Conclusion

IRFD, a predictive machine learning technique with unique feature engineering combination for detecting electricity theft is demonstrated in this study. To detect and remove outliers and evaluate the efficiency of our model, we have combined Recursive Feature Elimination with cross-validation (RFECV) and Isolation Forest (IF) in our proposed method as a novel approach. After that, Random Forest (RF) classifier is used in this case to predict electricity theft. Finally, the efficiency of our model is evaluated by comparing it to a variety of machine learning classifiers such as GBC, AdaBoost, KNN, LR, ANN, and XGBoost after and before feature engineering. The result section shows that before applying feature engineering, XGB has the best output with 94.2% accuracy. On the other hand, after applying feature engineering techniques, except LR, all other classifiers have surpassed their performance in terms of accuracy, precision, recall, and F1-score. The result also shows that RF has the best output with 97.06% accuracy after employing feature engineering methods.

The findings of this research could aid the power division in taking preventive measures to reduce non-technical losses such as electric theft. This experience can be replicated in the future by gathering large dataset with more features from power plants and invoices from the consumers in Bangladesh, as we have done with Tunisian data to find out and reduce the losses due to the manipulations of meters by consumers or for other circumstances, thus by helping evaluate their efficiency at the application level in this domain. Furthermore, a different combination of feature engineering methods with deep learning approaches can also be implemented.

References

[1] S. Aziz, S. Z. H. Naqvi, M. U. Khan, and T. Aslam, "Electricity theft detection using empirical mode decomposition and k-nearest neighbors," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*. IEEE, 2020, pp. 1–5.

[2] P. Sasirekha and R. Karthikeyan, "Non-technical loss detection in electric power distribution networks by using random forest fed support vector machines," *American International Journal of Research in Science, Technology, Engineering & Mathematics*, p. 96.

[3] M. Nabil, M. Ismail, M. M. E. A. Mahmoud, W. Alasmary, and E. Serpedin, "Ppetd: Privacy-preserving electricity theft detection scheme with load monitoring and billing for ami networks," *IEEE Access*, vol. 7, pp. 96 334–96 348, 2019.

[4] S. K. Singh, R. Bose, and A. Joshi, "Energy theft detection for ami using principal component analysis based reconstructed data," *IET Cyber-Physical Systems: Theory & Applications*, vol. 4, no. 2, pp. 179–185, 2019. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cps.2018.5050

[5] M. Ismail, M. F. Shaaban, M. Naidu, and E. Serpedin, "Deep learning detection of electricity theft cyber-attacks in renewable distributed generation," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3428–3437, 2020.

[6] S.-C. Yip, W.-N. Tan, C. Tan, M.-T. Gan, and K. Wong, "An anomaly detection framework for identifying energy theft and defective meters in smart grids," *International Journal of Electrical Power & Energy Systems*, vol. 101, pp. 189–203, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0142061517318719

[7] J. Ni, K. Zhang, K. Alharbi, X. Lin, N. Zhang, and X. S. Shen, "Differentially private smart metering with fault tolerance and range-based filtering," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2483–2493, 2017.

[8] S.-V. Oprea and A. Bâra, "Machine learning classification algorithms and anomaly detection in conventional meters and tunisian electricity consumption large datasets," *Computers & Electrical Engineering*, vol. 94, p. 107329, 2021.

[9] H. Huang, S. Liu, and K. Davis, "Energy theft detection via artificial neural networks," in *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*. IEEE, 2018, pp. 1–6.

[10] T. Ahmad, H. Chen, J. Wang, and Y. Guo, "Review of various modeling techniques for the detection of electricity theft in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 2916–2933, 2018.

[11] R. Razavi, A. Gharipour, M. Fleury, and I. J. Akpan, "A practical feature-engineering framework for electricity theft detection in smart grids," *Applied energy*, vol. 238, pp. 481–494, 2019.

[12] T. Hu, Q. Guo, X. Shen, H. Sun, R. Wu, and H. Xi, "Utilizing unlabeled data to detect electricity fraud in ami: A semisupervised deep learning approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3287–3299, 2019.

[13] A. Samoshyn, "Fraud detection in electricity and gas consumption," May 2020. [Online]. Available: https://www.kaggle.com/mrmorj/fraud-detection-in-electricity-and-gas-consumption

[14] M. U. Togbe, M. Barry, A. Boly, Y. Chabchoub, R. Chiky, J. Montiel, and V.-T. Tran, "Anomaly detection for data streams based on isolation forest using scikit-multiflow," in *International Conference on Computational Science and Its Applications*. Springer, 2020, pp. 15–30.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[16] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2009, pp. 475–482.

[17] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, 2015.