Research on Credit Card Default Prediction for Class-Imbalanced Datasets Based on Machine Learning

Jinyang Liu
School of Physical and Mathematical Sciences, Nanyang Technological University Singapore, 637371, Singapore

Keywords: Credit Card Default, Machine Learning, Class Imbalance Correction, Credit Risk, Classification Model.

Abstract: It's known that a robust credit relationship is advantageous for both parties involved. However, credit defaults significantly amplify risk for financial institutions. Hence, default rate prediction stands as a crucial objective for lending institutions and a well-functioning predictive model serves as a potent means to strengthen risk control. To this end, this paper constructed multiple machine learning classification models to achieve credit card default prediction. Feature selection, based on the importance of variables in the random forest, was implemented to enhance the model performance. The results shown that, addressing the skewed nature of credit default data, various SMOTE-based resampling methods were employed to improve data distribution and further optimize accuracy. Compared to other models, the random forest model demonstrated superior predictive effectiveness. After correcting the data distribution, there was a significant enhancement in the predictive performance of all models, with K-Means SMOTE showcasing outstanding performance in data correction and model accuracy optimization.

1 INTRODUCTION

As a convenient payment tool, credit cards have prospered steadily, becoming a pillar business for financial institutions and stoking the engine of consumption. However, this trend has also increased the challenges of risk management, as the probability of defaults expand, exposing banks to significant risk. According to a report from Wells Fargo, credit card delinquency rates are surging among commercial banks in 2023, with rates at smaller lenders even approaching 8%. This trend may foreshadow a future economic recession. Therefore, a well-performing default prediction model has become a vital focus of financial institutions. An accurate model can help institutions balance economic risks and returns, avoid overdue and bad debt, and ensure sustainable profits.

A large body of studies have shown great interest in credit card default prediction. Traditional credit risk measurement methods include discriminant analysis and logistic regression based on statistical principle (Yin et al 2013). In the past few years, researchers have gained fresh insights due to the significant potential offered by machine learning algorithms. Different models, including those founded on Logistic Regression, Decision Tree, and Random Forest, were detailed by Butare et al (Leo et al 2019). Through extensive and in-depth research, Zhou et al. introduced a predictive model designed for analyzing issues related to default (Butaru et al 2016). The empirical results show that decision tree is a fast and robust model to process high-dimension data with strong evaluation ability and high accuracy. In the study by Chen et al., the classification model combining k-means and BP neural network algorithms is formulated to training data and make prediction (Zhou et al 2019). Zeng et al. applied decision tree and random algorithms to establish credit card warning model respectively and concluded that the performance of random forest model is better (Chen and Zhang 2021). There is also a lot of research focused on exploring the strength of ensemble learning (Zeng et al 2020 & Kim et al 2019). Through a comparative analysis based on overdue customer payment data, Hamori suggested that boosting algorithms such as Random Forest, Boosting, and Neural Network outperform other conventional techniques (Ileberi, Sun and Wang 2021). Similarly, based on the analysis of financial data from small Chinese institutions, Zhu et al. considered that ensemble classifiers, which fully utilized the knowledge learned by multiple single

Liu, J.
Research on Credit Card Default Prediction for Class-Imbalanced Datasets Based on Machine Learning.
DOI: 10.5220/0012818100004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 441-447
ISBN: 978-989-758-690-3

classifiers, can effectively improve prediction accuracy (Hamori et al 2018).

However, a suitable model not only relies on advanced algorithms, but is also affected by the quality and distribution of the data. In the real world, bad credit customers make up only a small proportion of creditworthy customers. It has been shown that class-imbalanced data leads to deterioration in model performance (Zhu et al 2019). Therefore, effective data processing and feature selection techniques are necessary for skewed class scenarios (Abedin et al 2023). A large body of literature has addressed this problem. In skewed class classification, the most classical method is random resampling. However, random under-sampling has the potential to lose critical information about decision boundaries, while simple replication during random oversampling increases the likelihood of overfitting, both of which can affect the results of classification models (Alam et al 2020). More sophisticated heuristic approaches have been developed. A renowned technique proposed by Chawla et al., known as Synthetic Minority Over-sampling Technique (SMOTE), has found extensive applications across various domains (Chawla et al 2002). It generates new and reasonable samples for minority class based on the KNN algorithm. On this basis, Chen et al. proposed an improved algorithm that combines the SMOTE technique with the KNN algorithm -- K-Means SMOTE. Validation results demonstrate that this technique significantly enhances the predictive performance of the model (Chen and Zhang 2021). In addition, the borderline-SMOTE, which takes full account of the distributional characteristics of the samples, overcomes the limitations of boundary value processing and achieves the identification of noise and boundary samples (Han et al 2005).

Therefore, this study aims to construct multiple classification models and explore their application feasibility. Predictive models base on massive data and different artificial intelligence algorithms help decision-makers develop adaptive strategies to mitigate the adverse socio-economic impact of defaults. In addition, given the class-imbalanced nature of default data, various correction techniques are evaluated, so as to explore valid and effective processing methods to achieve further optimization of model prediction performance.

2 METHODOLOGY

2.1 Data Source

The dataset for the study is sourced from the Machine Learning Repositor of University of California, Irvine (UCI).

2.2 Data Explanation and Preliminary Analysis

The dataset contains 24 columns and 30,000 rows, of which about one-fifth of the samples are in the default category and the rest are in the non-default category. Table 1 demonstrates all the features of the dataset.

The default customer profile can be initially analysed by observing the data distribution of several key attributes. As shown in Figure 1, the majority of customers in arrears had relatively low limit balances, ranging from 20,000 to 50,000. As seen in Figure 2, 20–40-year-olds were the largest proportion of individuals in arrears in the dataset. Figures 3, 4 and 5 depict the gender, marital and educational status of clients. While there were slightly more females than males in the sample, females had slightly lower delinquency rates than males, while single individuals had slightly lower delinquency rates than married individuals. Credit card delinquency rates tended to decrease as educational attainment increased.

Table 1: Variable Attributes.

Number	Variable	Explanation
1	Limit_bal	Credit amount (NT$)
2	Sex	1 = male, 2 = female
3	Age	Years of age
4	Education	1= graduate school, 2 = university, 3 = high school, 4 = unknow
5	Marrige	1 = married, 2 = single, 3 = others
6 —11	pay_1 -- pay_6	Monthly disbursements April to September 2005
12 —17	bill_amt1 --bill_amt6	Amounts billed from April to September 2005
18 —23	pay_amt1 — pay_amt6	Prior Period Payment Amount from April to September 2005 (NT$)
24	Default.payment.next.month	A binary variable, as the response variable. Yes = 1, No = 0

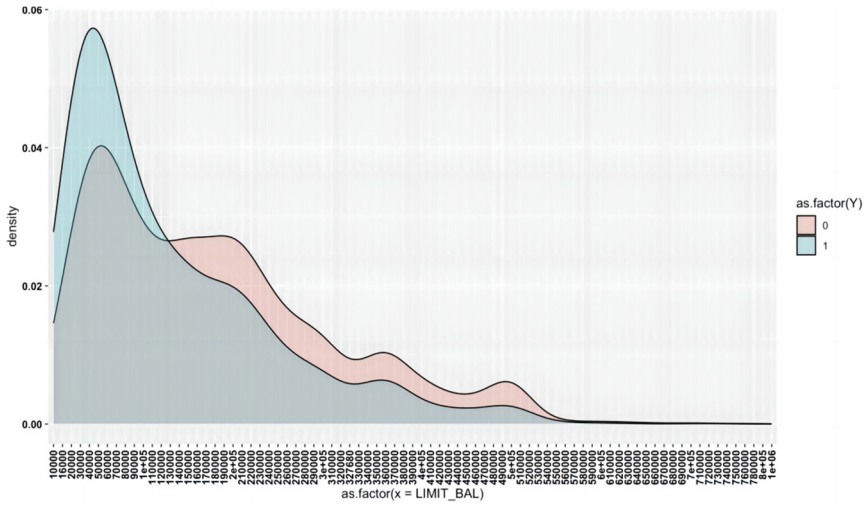

Figure 1: Density diagram of LIMIT_BAL (Picture credit: Original).

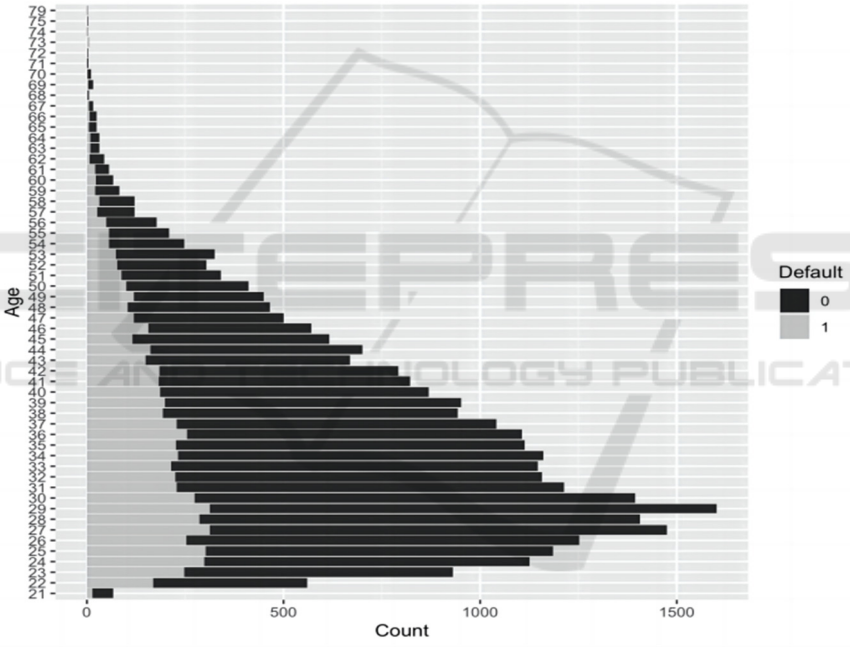

Figure 2: Density diagram of AGE (Picture credit: Original).

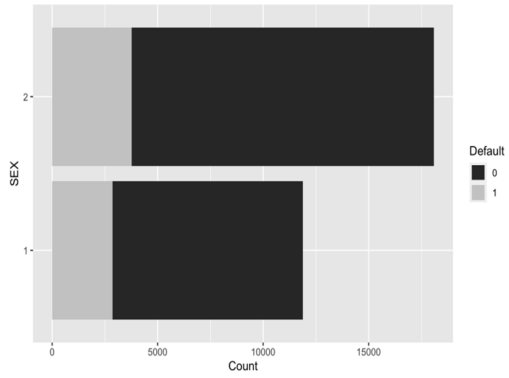

Figure 3: Bar chart of SEX (Picture credit: Original).

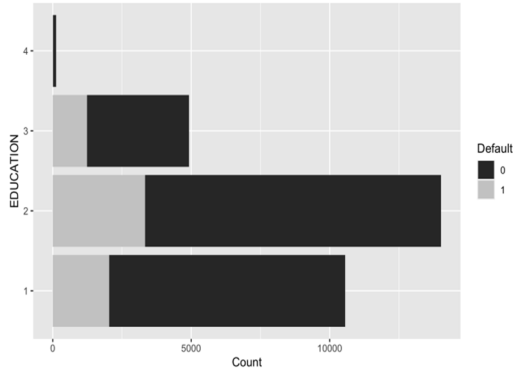

Figure 4: Bar chart of Education (Picture credit: Original).

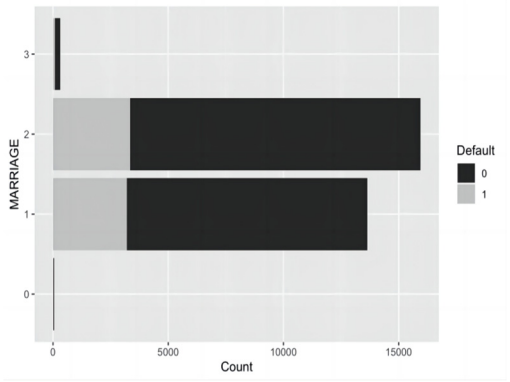

Figure 5: Bar chart of Marriage (Picture credit: Original).

2.3 Modelling Techniques

2.3.1 Logistic Regression

The Logistic Regression is a linear model that maps the linear combination of features into the logistic function, transforming the real values into a range between 0 and 1, representing the probability of belonging to a particular class.

2.3.2 K-Nearest Neighbors

The K-Nearest Neighbors algorithm (KNN) is a non-parametric supervised learning method widely applied in pattern recognition and classification tasks. Based on the principles of the algorithm, the category of an object is not solely determined by the majority vote of its neighbors but also involves considerations of distance weights for each neighbor. This approach excels in leveraging the local structural information among samples, proving effective across a diverse range of data types.

2.3.3 Decision Tree

A decision tree is a tree model that makes predictions by splitting data into subsets based on features. The goal is to select the best features to split at each node with the aim of maximising information gain (for classification) or minimising variance (for regression). It can model complex non-linear data relationships and also handle numerical and categorical features.

2.3.4 Random Forest

Random forest is an integrated learning method, a special form of Bagging, that makes predictions based on a collection of decision trees. In constructing each decision tree, random forest employs bootstrap and random feature selection to increase the diversity of the model. In this study, the final prediction was obtained by majority voting.

2.4 Class Imbalance Correction Techniques

Class imbalance creates significant challenges to the results of most classification algorithms. This emanates from the inherent constraint imposed on the model's learning and analytical capabilities due to uneven data distribution. Hence, three resampling techniques were employed to investigate their efficacy in handling skewed data and assess the extent to which they could optimise model performance.

Synthetic minority oversampling technique (SMOTE), is based on the KNN algorithm, which measures the characteristics of the Kth nearest neighbour of a particular sample and calculates the characteristics to create a new sample based on the degree of difference (Chawla et al 2002). Borderline-SMOTE is an improved extension of the original SMOTE algorithm, whose goal is to improve the performance of the classifier by identifying samples located at the border to generate new synthetic samples (Han et al 2005). K-Means SMOTE is a combination of K-Means clustering and SMOTE, which clusters the sample data and filter clusters with more minority categories for SMOTE oversampling (Chen and Zhang 2021). This algorithm can reduce the imbalance both between and within categories.

3 RESULTS AND DISCUSSION

3.1 Initial Model Prediction

Following fundamental steps of data preprocessing and normalization, multiple classification models are constructed base on Five-Fold Cross-Validation. The results of the model evaluation are shown in Table 2.

It can be observed from Figure 6 that among the four models, the performance of the Random Forest is relatively superior, with most evaluation metric values distinctly surpassing those of the other models. However, it is noteworthy that, despite the accuracy of all four models not being excessively low—except for Logistic Regression, where the accuracy of the other three models exceeds 70%—the metrics such as MCC, Precision, Recall, and F1 score are relatively low, mostly falling below 0.5. This indicates a diminished predictive capability of the models for classifying the minority class in this scenario.

3.2 Model Prediction After Feature Engineering

Feature engineering constitutes an effective technique aimed at diminishing model complexity and alleviating noise interference. In this study, feature selection relies on the variable importance calculated by random forest model, using the Gini coefficient as a metric to gauge the contribution of each feature. As shown in Table 3 and Figure 7, there is no substantial improvement in various metrics, with only marginal fluctuations.

Table 2: Comparison results of four models.

Model	Accuracy	MCC	Precision	Recall	F1 score
Logistic	0.611	0.174	0.306	0.584	0.393
KNN	0.755	0.135	0.386	0.182	0.247
Decision Tree	0.729	0.216	0.388	0.391	0.390
Random Forest	0.815	0.379	0.655	0.343	0.450

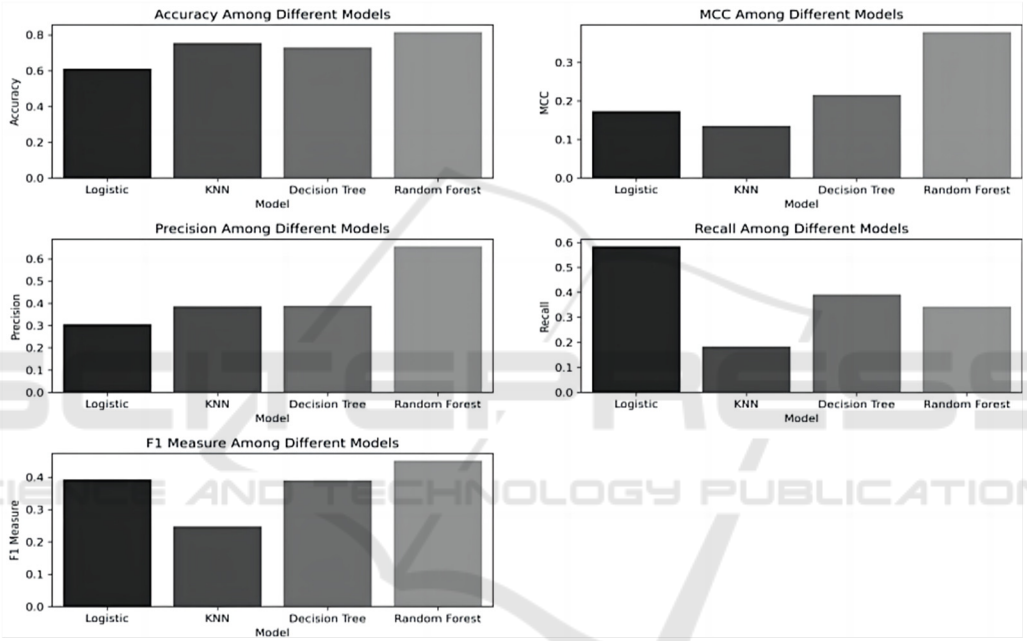

Figure 6: Comparative evaluation of models (Picture credit: Original).

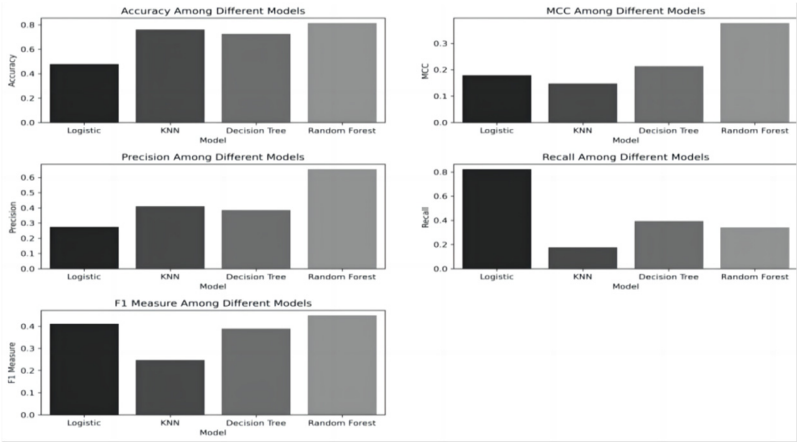

Figure 7: Comparative evaluation after feature engineering (Picture credit: Original).

Table 3: Model comparison after feature engineering.

Model	Accuracy	MCC	Precision	Recall	F1 Score
Logistic	0.479	0.179	0.274	0.822	0.411
KNN	0.762	0.148	0.410	0.177	0.247
Decision Tree	0.726	0.213	0.385	0.394	0.389
Random Forest	0.814	0.377	0.654	0.342	0.449

Table 4: Model comparison after oversampling.

Oversampling Methods	Performance Measure	Logistic	KNN	Decision Tree	Random Forest
SMOTE	Accuracy	0.601	0.746	0.814	0.884
	MCC	0.211	0.509	0.628	0.771
	Precision	0.637	0.696	0.806	0.922
	Recall	0.473	0.873	0.827	0.839
	F1 score	0.535	0.535	0.816	0.879
Borderline-SMOTE	Accuracy	0.588	0.755	0.813	0.882
	MCC	0.185	0.528	0.626	0.768
	Precision	0.627	0.702	0.806	0.920
	Recall	0.437	0.887	0.823	0.838
	F1 score	0.511	0.783	0.815	0.877
KMeans-SMOTE	Accuracy	0.722	0.810	0.820	0.886
	MCC	0.450	0.624	0.640	0.775
	Precision	0.718	0.782	0.813	0.922
	Recall	0.737	0.861	0.832	0.844
	F1 score	0.725	0.819	0.822	0.881

3.3 Model Prediction after Handling Class Imbalance

Table 4 below illustrates the updated performance metrics of all models after applying three resampling techniques to address the class imbalance problem. As shown in the table, there is no significant change in the accuracy rates. This is primarily attributed to the skewed distribution of the majority class in the imbalanced data, which tends to keep the accuracy consistently at a high level. However, it is evident that, after correcting the data distribution, we have successfully narrowed fown the errors of the mosels across different classes, enabling them to accurately capture complex patterns present in the real world. This process has brought substantial benefits to the performance enhancement of the models, especially the Precision, Recall, and F1 scores of the Decision Tree and Random Forest models exceeded 0.8, which were even less than 0.4 before the class correction. This suggests a substantial enhancement in the predictive accuracy of the models for the minority class, laying a solid groundwork for further practical applications.

Among the three correction techniques, K-Means SMOTE exhibits the most favorable performance, achieving a greater degree of optimization in model predictive performance while addressing data distribution imbalances.

3.4 Discussion

Before addressing the class imbalance problem, although the prediction accuracies of all four models exceed 60%, the models' prediction performance for minority class (defaulting customers) in skewed categorization is not satisfactory. The main reason is that it is misleading to focus solely on the precision rate in the case of a severe imbalance of positive and negative category samples. It is necessary to establish comprehensive "unbiased" evaluation metrics, such as Precision, Recall, and MCC (Matthews correlation coefficient), to avoid the partiality of a single metric and enhance a holistic understanding of model performance (Boughorbel et al 2017).

After correcting the data distribution using the resampling technique, the model’s performance significantly improved, especially the prediction accuracy for the minority class. Among the three methods, KMeans-SMOTE had a more positive impact on the model’s performance, possibly because KMeans-SMOTE is better at generating synthetic samples with intrinsic distributional characteristics after clustering (Chen and Zhang 2021).

Among all classification models, the superiority of random forest has been thoroughly validated. The ensemble algorithm combines ideas by aggregating multiple weak classifiers and introduces randomness to prevent the model from overly relying on specific

data, thereby improving its generalization performance when faced with real-world complex data. However, a number of studies have shown that determining a universally effective model that performs well across a majority of institutions is challenging due to variability in customer information and credit metrics (Butaru et al 2016). Therefore, the final model selection and data processing should depend on the dataset characteristics, sample distribution, and performance requirements. In future research, it is recommended to explore various variants of traditional algorithms for further enhancement. Simultaneously, expanding the scope of the study by including a broader range of economic indicators and user behavioural metrics into the training data, it is more likely to establish a model that integrates multiple perspectives, thoroughly considers data diversity.

4 CONCLUSION

In summary, based on machine learning principles, this study constructed classification models including logistic regression, KNN, decision trees and random forests for predicting credit card default. This study also compared various SMOTE-based resampling techniques to correct the data distribution and evaluate their impact on improving the performance of predictive models. From the results, the model based on the random forest algorithm had higher generalisability and prediction accuracy. The research also indicates that dealing with class imbalance data can significantly enhance the prediction accuracy for minority categories while maintain robustness for majority groups. Therefore, the key to building effective default prediction models lies in the use of sound and superior algorithms combined with efficient data processing methods. Future explorations can further delve into more advanced algorithms and techniques to uncover more robust results. This will help financial institutions construct a comprehensive credit risk prediction and credit assessment system to lower financial risk. With the continuous strengthening of financial regulations, default prediction is poised to become a vital tool in risk management, offering financial institutions judgment criteria and decision support, thereby fostering the stable operation and healthy development of financial markets.

REFERENCES

L. Yin, Y. Ge, K. Xiao, X. Wang, X. Quan, Neurocomputing, 105, 3-11 (2013).
M. Leo, S. Sharma, K. Maddulety, Risks, 7(1), 29 (2019).
F. Butaru, Q. Chen, B. Clark, et al.，J. Banking & Finance, 72, 218-239 (2016).
J. Zhou, W. Li, J. Wang, S. Ding, C. Xia, Physica A: Stat. Mech. Appli., 534, 122370 (2019).
Y. Chen, R. Zhang, Complex, 1-13 (2021).
X. Zeng, L. Lu, X. Lu et al. Wireless Internet Tech., 17(18), 166-168 (2020).
E. Kim, J. Lee, H. Shin, et al., Expert Syst Appl, 128, 214-224 (2019).
E. Ileberi, Y. Sun, Z. Wang, IEEE, 9, 16528-165294 (2021).
S. Hamori, M. Kawai, T. Kume, Y. Murakami, C. Watanabe, J. Risk Finan. Manage., 11(1), 12 (2018).
Y. Zhu, L. Zhou, C. Xie, G.J Wang, T.V. Nguyen, Int J Prod Econ, 211, 22-33 (2019).
M. Z. Abedin, C. Guotai, P. Hajek, T. Zhang, Comp. Intel. Sys., 9(4), 3559-3579 (2023).
T.M. Alam, K. Shaukat, I.A. Hameed, et al., IEEE, 8, (2020).
N. Chawla, K. Bowyer, L. Hall, J. Arti. Intel. Res., 16, 321-357 (2002).
H. Han, W. Y. Wang, B. H. Mao, Inter. Conf. Intel. Comp., Berlin, Heidelberg: Springer Berlin Heidelberg, (2005).
S. Boughorbel, F. Jarray, M. El-Anbari, PloS one, 12(6), (2017).