Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Improving Bankruptcy Prediction Using Machine Learning

T Sai Lalith Prasad[1], K Neeraja Reddy[2], G Eeshita[2], K Sai Shivani[2]

[1]Assistant Professor, Department of Artificial Intelligence and Data Science, Vignan Institute of Technology and Science, Hyderabad, India

[2]UG Student, Department of AI&DS, Vignan Institute of Technology and Science, Hyderabad, India

Correspondence

T. Sai Lalith Prasad

Assistant Professor , Department of Artificial Intelligence and Data Science, Vignan Institute of Technology and Science, Hyderabad, India

Abstract

It is very crucial to predict bankruptcy for the decision-making process of creditors, investors, businesses, and policymakers. While it will help businesses and financial institutions in making sound judgments by accurately projecting bankruptcy, it helps to reduce the adverse impacts also for the economy and society. The methodologies like random forest regression and SMOTE along with other algorithms were cross validated to enhance accuracy. These prediction models can be developed further by including both financial and non-financial factors like market conditions and management quality. Moreover, dramatic efficiency improvements can be achieved through advanced technologies like deep learning. Technological advancement and data accessibility will increase these tactics such that banks and businesses can discover bankruptcy risk, make the right decisions, and save losses.

- Received Date: 25 May 2025
- Accepted Date: 15 June 2025
- Publication Date: 27 June 2025

Keywords

bankruptcy prediction, machine learning, SMOTE, random forest classification, cross-validation, accuracy.

Introduction

The improvement of BPMs amplifies past conventional factual strategies since all machine learning and fake insights calculations have presently ended up a necessarily portion in developing them. Other strategies like outfit learning, neural systems, and profound learning strategies give way better exactness since complex, nonlinear connections in the information set are subjected to point by point investigation. These methods can be connected to distinctive sorts of information. They can be connected to well-organized monetary information and to unorganized information, such as news articles, social media sentiments, or expansive financial measures. One of the major issues that BPMs confront is managing imbalanced datasets. In these datasets, there are exceptionally few bankrupt companies compared to non-bankrupt ones. Procedures such as Destroyed (Engineered Minority Oversampling Strategy) can be utilized by making manufactured tests for the minority classes to progress demonstrate execution. In expansion, including building and determination are vital in making beyond any doubt that as it were the most pertinent factors contribute to expectations to minimize clamor and make strides exactness. Ensuing advancement of Commerce Handle Administration frameworks will depend on huge information analytics since comprehensive datasets obtained from different sources can offer an organization a more in-depth understanding of its budgetary circumstance. In expansion, integration with cloud computing and edge advances will improve the capacity to prepare real-time information so that chance assessments may be made in an opportune way. The integration of Commerce Handle Administration into decision-support systems will increment the proficiency of bank loaning and venture methods as much as that of businesses themselves. Moral considerations-such as straightforwardness and fairness-are especially imperative in the plan of the Commerce Prepare Administration frameworks that will guarantee that inclinations built into information or calculations lead to out of line results. Rules and administrative systems will set the guidelines for these models, guaranteeing validity and responsibility. In conclusion, the ceaseless change of liquidation forecast models will not be as if helping partners make superior choices but to guarantee budgetary soundness and versatility in the by and large economy. Liquidation forecast models can significantly diminish the unfavorable impacts of corporate disappointments through the integration of inventive procedures, comprehensive datasets, and moral hones. Liquidation emerges when it is accepted that an enterprise can no longer pay off all its extraordinary obligations. In such a situation, grave challenges emerge for companies in expansion to banks. As the number of trade disappointments increments, so does the requirement for viable BP models.

Citation: Prasad TSL, Reddy KN, Eeshita G, Shivani KS. Improving Bankruptcy Prediction Using Machine Learning. GJEIIR. 2025;5(4):072.

Hence, such models are instrumental apparatuses in the appraisal of any company's budgetary solidness as a premise for expanding credits, advances, or performing contracts. An early discovery of monetary trouble from a BPM goes a long way in anticipating trade closure dangers and contributes to successful bank, financial specialist, and policy-maker choices. The model's rightness and execution basically lie with the quality, volume, and differences of information and calculations conveyed whereas building such models. Over time, adjust sheets and salary explanations had been conventional inputs into building BPMs. But current patterns appear that joining components such as the quality of administration, showcase patterns, and the circumstance in the industry may include exactness in anticipating. The strategies of machine learning incorporate irregular timberland relapse and destroyed, which have been known to perform well with huge sums of complicated information and issues with a lesson awkwardness. Profound models moreover found little designs not distinguished by other strategies. As information gets to be progressively open and computational innovation creates encouragement, BPMs will proceed to become much more dependable, proficient, and flexible. This may in this manner upgrade the capacity of businesses and monetary teach to overcome dangers and guarantee soundness; encourage minimize misfortunes; and accomplish a financially more grounded environment overall.

Methodology

The data is collected from the Taiwan Economic Journal, 1999-2009.The dataset had 6,819 cases with 96 features like financial ratios and performance measures. It also had a binary target showing if a company was bankrupt (1) or not (0).Once data was put into a Panda DataFrame, duplicate and null values were verified, and none were found. The Synthetic Minority Oversampling Technique (SMOTE) was applied to bring data balance in the dataset by creating artificial samples for the minority class, and exploratory data analysis was conducted to bring out the underlying patterns with respect to the relationship of the target variable and its attributes to visualize, and Select KBest was applied for feature selection. The dataset was stratified into the training and testing subsets at an 80-20% ratio. Stratified K-Fold cross-validation was adopted to ensure proper stratification of the data. For model development, we ran Random Forest, KNN, SVC, Logistic Regression, and Decision Tree in addition to using Randomized Search Cross-Validation on the hyperparameter tuning. The models are tested using accuracy, precision, recall, F1 score, and ROC-AUC.

Modelling and analysis

In the modeling phase, the dataset was divided into training and testing subsets, allocating 80% of the data for training purposes and reserving 20% for testing. This division guarantees that the model is developed using a substantial segment of the data while still maintaining enough for subsequent evaluation. Due to the huge imbalance of the current dataset, a stratified K-Fold cross-validation was used to ensure that there is an adequate number of both classes during the training and test partitions. Several machine learning algorithms are used to determine the most effective one in bankruptcy prediction. These models that are used for the implementation of the systems include Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, and Decision Tree. All the algorithms were trained and tested, and their effectiveness was measured in terms of accuracy, precision, recall, F1 score, and ROC-AUC. The Random Forest classifier emerged as the best performer, achieving an accuracy of 98.9% and an F1 score of 0.86. This model uses an ensemble of decision trees to make predictions, which makes it more generalizable and less prone to overfitting. The KNN algorithm performed well too, with 98.02% accuracy and an F1 score of 0.78. However, it's slower for larger datasets since it computes the distance between points at predicted time. The SVC algorithm achieved an accuracy of 94.43%, but its F1 score was lower at 0.56, which shows it had difficulty with imbalanced data. The classifiers based on Logistic Regression and Decision Trees had lower accuracies at 88.86% and 84.68% respectively, and their F1 scores were also less desirable than that of the top models. This reflects that although models like Logistic Regression are simpler and possess higher interpretability, they fail to achieve the performance that ensemble methods like Random Forest can attain on complex datasets. I have used SelectKBest technique for feature selection to obtain the best features that are most important and would therefore help in making good predictions. It does not reveal any enhancements in the models' performances since the features in this dataset are too numerous and therefore relevant. Overall, the results demonstrated that Random Forest and KNN are the most accurate models for predicting the bankruptcy of a company. It was also found that the top choice is Random Forest due to its high accuracy and F1 score. The models are tested using accuracy, precision, recall, F1 score, and ROC-AUC.

Architecture description

The image depicts a predefined pipeline for predicting bankruptcy in a company, using machine learning (ML)

Dataset

- The pipeline starts with a dataset that contains financial, as well as operational data on companies; it contains features that could affect bankruptcy outcomes, such as revenue, shortlisted for debt levels, and other financial ratios.

Data Preprocessing

- Raw data are normally contaminated with noise, contain null values, or exhibit outlier features. This stage cleans the data by addressing missing values, normalizing features, and putting data in a format that is easy to analyze.

Prepared Dataset

- Once its preprocessing is finalized, data ready to be processed are termed a "prepared dataset"; further pursuance of intention leads to splitting this prepared dataset into two parts:

- Training Set: in other words, configured to build and tune the ML model

- Testing Set: wherein test models get evaluated on unseen data.

K-fold Cross-Validation

- It is employed in the efficient process of training. The training set is generally divided into several subsets (folds), and each fold is used for validation after the model has been trained on the other folds. This guarantees the model's ability to generalize effectively while avoiding overfitting.

Oversampling Module

- Bankruptcy data sets in general are often imbalanced because the number of bankrupt companies is way less than that of non-bankrupt ones. In order to prevent the model from being biased toward the majority class, oversampling methods, such as SMOTE (Synthetic Minority Over-sampling Technique), are applied to the training set to balance the classes.

Training

- A well-prepared and balanced training set for training the algorithms of ML. Various models can be adopted here, elements include logistic regression, random forests, or neural networks.

Machine Learning Algorithms

- These refer to the algorithms that run at the core of computational models used for pattern learning by relying on the data available. Based on learning, the algorithms yield the results and provide predictions regarding the possibility of a company going bankrupt.

Testing

- Once the phase of training of the model has been finished, it is evaluated using a separate testing set to provide insight into its performance on unknown data.

Performance evaluation

- The model's prediction is assessed with the performance measures accuracy, precision, recall, F1-score, and ROC-AUC to evaluate the model's performance in forecasting bankruptcy.

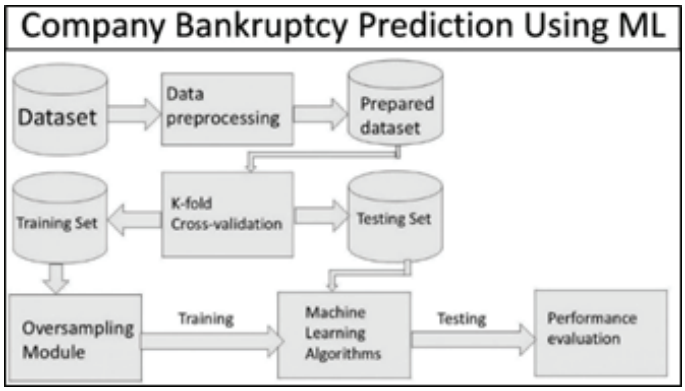


Figure 1. *Architectural Design*

Importing modules

Pandas and NumPy

- The library most associated with data manipulation and analysis is pandas. Importing a dataset into a DataFrame, cleaning data, and then performing tasks such as filtering and grouping is a great example of how the process works.
- NumPy supports numerical computations and the execution of array operations that are of great importance for pre- processing data and feature construction.

Matplotlib and Seaborn

- Matplotlib is basically used to make simple plots and graphics like line plot, bar charts, histogram.
- Seaborn is built on top of Matplotlib and is used for creating advanced graphics such as heatmaps, pair plots and other plots that are extremely useful during EDA.

Scikit-learn

- Provides functionality to divide a dataset into the training set and the testing set, with other preprocess functionalities including scaling features amongst others. Other closely related techniques encompass all prominent ML algorithms including Random Forest, KNN, SVC, Logistic Regression and Decision Tree.
- This package has cross-validation, optimization of hyperparameters and even feature selection

SMOTE (Synthetic Minority Oversampling Technique)

- SMOTE is the algorithm in the imbalanced-learn library for handling class imbalance through creating new samples for the minority class.

Warnings:

- Warnings are imported from the Warnings library to eliminate the warnings produced during the execution and training of a model so that outputs are less messy.

```
import numpy as np
import pandas as pd
import seaborn as sns
from random import randint
import matplotlib.pyplot as plt

from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import train_test_split
from sklearn.model_selection import StratifiedKFold
from imblearn.pipeline import make_pipeline as imbalanced_make_pipeline
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import classification_report, confusion_matrix, f1_score,accuracy_score, precision_score, recall_score, roc_auc_score

from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import f_classif

import warnings
warnings.filterwarnings("ignore")
```

Figure 2. *Importing Libraries*

Output of Working Model

In the exploratory data analysis, visualizations were quite useful in identifying significant patterns. The analysis showed that only a few firms went bankrupt when they had more assets than liabilities. Factors such as the Debt Ratio, Current Liabilities to Assets Ratio, and Current Liabilities to Current Assets Ratio were found to have strong correlations with bankruptcy events. Firms with higher levels of assets and revenue had a lesser chance of bankruptcy, which was indicated by the negative correlations found in these attributes.

The study tested several machine learning algorithms to predict corporate bankruptcies. Among the suite of models used, it is the Random Forest Classifier that performed better than the others as this reached an accuracy level of 98.9%, with an F1 of 0.86; thus, higher accuracy and recall indicate that this model is most reliable for bankruptcy predictions involving companies.

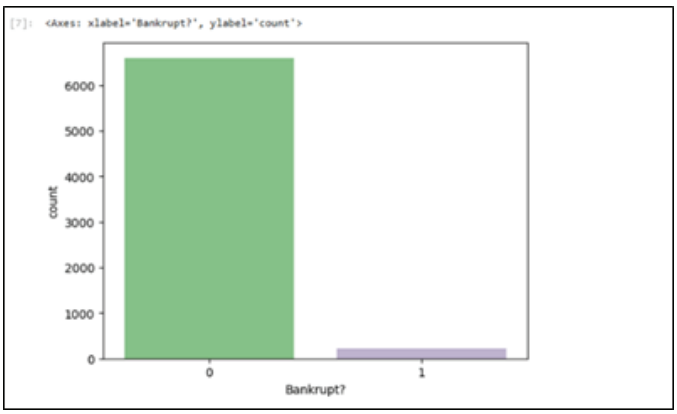


Figure 3. *Importing Libraries*

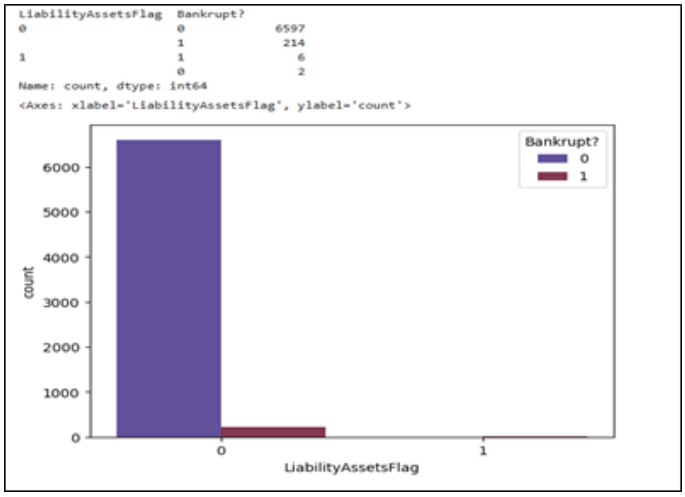


Figure 4. *Finding correlation between bankrupt & liability assets flag features*

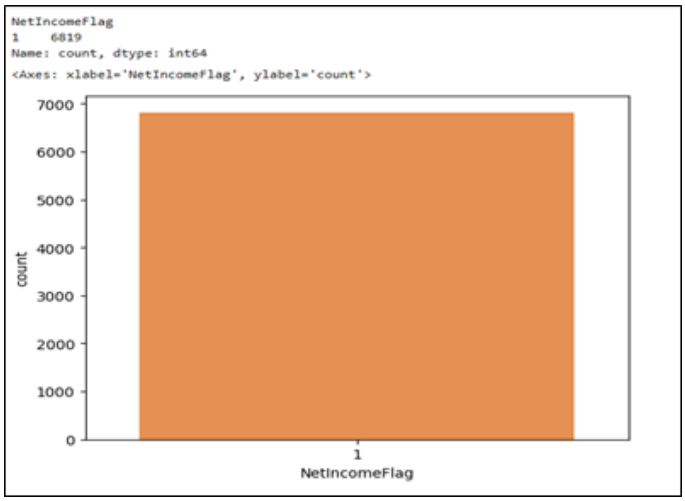


Figure 5. *Counting total number of net income flag*

The K-Nearest Neighbors algorithm was found to be effective with an accuracy of 98.02% and an F1 score of 0.78.

The accuracy of alternative models, like Support Vector Classifier (SVC), Logistic Regression, and Decision Tree, decreased to 94.43%, 88.86%, and 84.68%, respectively. Their F1 scores are also almost on the same track. This means that even though models like Logistic Regression provide interpretability, more complex models, like Random Forest, better fit this data set as they perform well.

Conclusion

The research studied the dataset, developed the forecasting models, and estimated the possibility of corporate bankruptcy of a firm. A preliminary analysis of the data probed into the major trends and the association between financial variables and bankruptcies. The results show that companies with a higher "Debt Ratio %," "Current Liability to Assets," and "Current Liability to Current Assets" have a higher chance of bankruptcy. In contrast, firms with more assets and better earnings are less likely to go bankrupt.

The best models were the Random Forest Classifier and K-Nearest Neighbors (KNN), with Random Forest recording

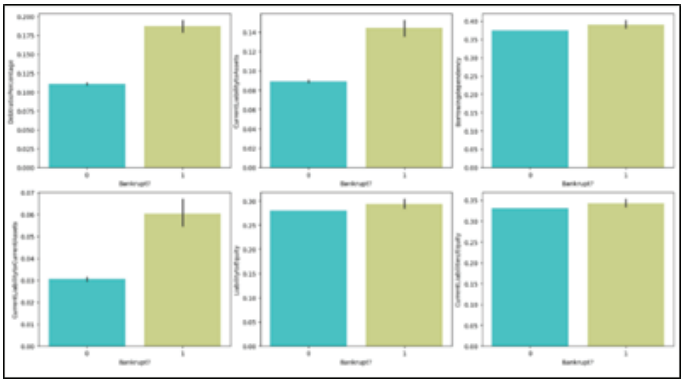
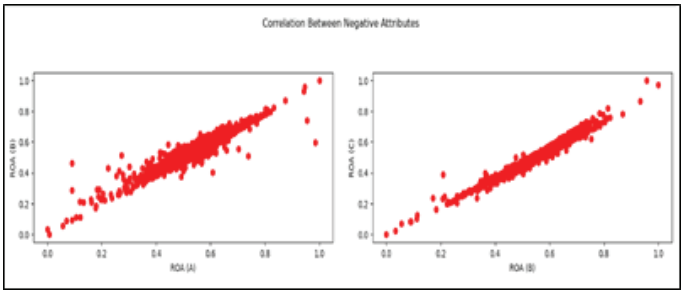


Figure 6. *Analysing top six positively correlated attributes*



[25]:

	Algorithm	Model Score	Precision	Recall	F1 score	ROC-AUC score
0	K Nearest Neighbour	98.02%	0.67	0.94	0.78	0.96
4	Random Forest Classifier	96.04%	0.49	0.98	0.65	0.97
3	Support Vector Classifier	95.82%	0.47	0.90	0.62	0.93
1	Logistic Regression	89.22%	0.23	0.82	0.36	0.86
2	Decision Tree Classifier	84.68%	0.18	0.90	0.31	0.87

Figure 8. *Performance of the model chosen*

29]:

	Algorithm	Model Score	Precision	Recall	F1 score	ROC-AUC score
0	Random Forest Classifier	98.83%	0.81	0.90	0.85	0.95
1	K Nearest Neighbour	97.8%	0.64	0.94	0.76	0.96

Figure 9. *Best performing models*

the peak accuracy of 98.9% and an F1 score of 0.86. These results testify to the strength of these models in bankruptcy prediction. Through this study, the task of developing a model of high accuracy in predicting corporate bankruptcy was achieved, which then significantly contributes to the overall financial decision-making processes involved.

Future enhancements

This project presents very significant opportunities for improvement and extension in many ways. For example, it can be made more robust by including additional years of financial data or even adding more country data. Also, different types of data can be added to this research, such as market trends or management practice by companies, which might provide more

fruitful insights in predicting bankruptcy. This would mean that other methods such as developing new features or applying PCA techniques will be useful in determining the hidden patterns, hence making the data more understandable. The project can eventually be created as a real-time web application where users may input financial data and then receive predictions shortly after the input. Additionally, it could learn dynamically, meaning that its model would keep updating with more data, as they continue to arise, to continue being useful in the real world and to remain correct.This will make the project more useful and connected to real life.More complex algorithms such as XGBoost or LightGBM can contribute to a significant difference in the model's performance. Ensemble methods that use many models combined also increase accuracy significantly. Tools like SHAP or LIME may be very helpful in making model predictions much more understandable to humans..

References

1. “IMPROVING BANKRUPTCY PREDICTION USING MACHINE LEARNING” by Tran Duc Quynh, Tran Thi Lan Phuong.https://ieeexplore.ieee.org/document/9287707/
2. Towards Data ScienceComprehensive articles on machine learning, data preprocessing, and feature selection. https://towardsdatascience.com
3. GeeksforGeeksTutorials and examples on Random Forest, SMOTE, and other machine learning techniques. https://www.geeksforgeeks.org
4. Analytics VidhyaGuides and insights on building predictive models and using evaluation metrics.https://www.analyticsvidhya.com
5. Scikit-learn DocumentationOfficial documentation for implementing machine learning algorithms and tools. https://scikit-learn.org
6. R. Bhallamudi et al., "Deep Learning Model for Resolution Enhancement of Biomedical Images for Biometrics," in Generative Artificial Intelligence for Biomedical and Smart Health Informatics, Wiley Online Library, pp. 321–341, 2025.
7. R. Bhallamudi et al., "Artificial Intelligence Probabilities Scheme for Disease Prevention Data Set Construction in Intelligent Smart Healthcare Scenario," SLAS Technology, vol. 29, pp. 2472–6303, 2024, Elsevier.
8. R. Bhallamudi, "Improved Selection Method for Evolutionary Artificial Neural Network Design," Pakistan Heart Journal, vol. 56, pp. 985–992, 2023.
9. R. Bhallamudi et al., "Time and Statistical Complexity of Proposed Evolutionary Algorithm in Artificial Neural Networks," Pakistan Heart Journal, vol. 56, pp. 1014–1019, 2023.
10. R. Krishna et al., "Smart Governance in Public Agencies Using Big Data," The International Journal of Analytical and Experimental Modal Analysis (IJAEMA), vol. 7, pp. 1082–1095, 2020.
11. N. M. Krishna, "Object Detection and Tracking Using YOLO," in 3rd International Conference on Inventive Research in Computing Applications (ICIRCA-2021), IEEE, Sept. 2021, ISBN: 978-0-7381-4627-0.
12. N. M. Krishna, "Deep Learning Convolutional Neural Network (CNN) with Gaussian Mixture Model for Predicting Pancreatic Cancer," Springer US, vol. 1380-7501, pp. 1–15, Feb. 2019.
13. N. M. Krishna, "Emotion Recognition Using Skew Gaussian Mixture Model for Brain–Computer Interaction," in SCDA-2018, Textbook Chapter, ISBN: 978-981-13-0514, pp. 297–305, Springer, 2018.
14. N. M. Krishna, "A Novel Approach for Effective Emotion Recognition Using Double Truncated Gaussian Mixture Model and EEG," I.J. Intelligent Systems and Applications, vol. 6, pp. 33–42, 2017.
15. N. M. Krishna, "Object Detection and Tracking Using YOLO," in 3rd International Conference on Inventive Research in Computing Applications (ICIRCA-2021), IEEE, Sept. 2021, ISBN: 978-0-7381-4627-0.
16. T. S. L. Prasad, K. B. Manikandan, and J. Vinoj, "Shielding NLP Systems: An In-depth Survey on Advanced AI Techniques for Adversarial Attack Detection in Cyber Security," in 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), IEEE, 2024.
17. S. Sowjanya et al., "Bioacoustics Signal Authentication for E-Medical Records Using Blockchain," in 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), vol. 1, IEEE, 2024.
18. N. V. N. Sowjanya, G. Swetha, and T. S. L. Prasad, "AI Based Improved Vehicle Detection and Classification in Patterns Using Deep Learning," in Disruptive Technologies in Computing and Communication Systems: Proceedings of the 1st International Conference on Disruptive Technologies in Computing and Communication Systems, CRC Press, 2024.
19. C. V. P. Krishna and T. S. L. Prasad, "Weapon Detection Using Deep Learning," Journal of Optoelectronics Laser, vol. 41, no. 7, pp. 557–567, 2022.
20. T. S. L. Prasad et al., "Deep Learning Based Crowd Counting Using Image and Video," 2024