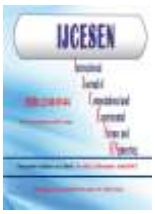
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 5790-5798
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Enhancing Trade Balance Prediction in Iraq Using Optimized Random Forests with Synthetic Data Augmentation

Asmaa Abdul Azeez Dakhil[1]*, Ali Hasan Taresh[2]

[1]Information Technology & Communications University Baghdad, Iraq
* **Corresponding Author Email:** Ms202320745@iips.edu.iq - **ORCID:** 0000-0002-5247-1850

[2]Information Technology & Communications University Baghdad, Iraq
Email: alihtaresh@uoitc.edu.iq - **ORCID:** 0000-0002-8776-0322

Article Info:

DOI: 10.22399/ijcesen.3615
Received : 25 May 2025
Accepted : 30 July 2025

Keywords

Trade balance,
Random forest,
Synthetic data generation,
GMM,
hyperparameter optimization,
economic forecasting

Abstract:

This research applied the Random Forest algorithm to analyze and forecast the trade balance based on a set of binary economic indicators (surplus/deficit) to meet classification requirements. The trade balance is a critical economic indicator that represents the difference between a country's exports and imports. Several techniques were used to enhance the prediction accuracy. One of these techniques is generating synthetic data using generalized matrix models (GMMs) to amplify the data, which helped the model avoid overfitting. A set of parameters of the Random Forest algorithm were also modified to enhance the prediction accuracy. The model was trained and evaluated 30 times using random variables generated by the random number function to ensure its stability, achieving an accuracy of 98.23%. These results explains the effectiveness of machine learning algorithms in processing economic data and extracting patterns from them.

1. Introduction

Nuraini,2019 focused on the general concept of the trade balance, which measures a country's exports and imports. Whenever exports exceed imports, a country's trade balance is in surplus, while conversely, when exports are lower, the trade balance is in deficit. The researcher also focused on the factors that can affect the trade balance[1].

Saleh , 2022 focused his research on the Iraqi economy's near-total dependence on crude oil exports, which accounted for more than 99% of total export value, leading to the collapse of productive sectors and increased dependence on imports. [2].

Alkooranee et al., 2023Their study examined the impact of economic reform policies on Iraq's trade balance. The policies revealed a trade deficit due to reliance on oil and weak productive sectors. The study used the ARDL model, bounds testing, and an error correction model to measure the relationship between economic growth and the trade balance. [3].

Marie et al., 2023 Their study analyzed the impact of public debt on the trade balance in Iraq, showing that the expansion in public debt did not contribute to improving the trade balance. The ARDL regression model and the bound test were used. [4].

Al-Saeedi's ,2022 study focused on the port of Faw, the dry canal, job opportunities, and the profits that will be available, explaining that Iraq will have an important position on the global trade map [5].

Abellanis et al. (2024) demonstrated another method for generating synthetic data using Bayesian Gaussian (BGM). The results obtained after data generation were characterized by high accuracy and similarity to the original data, especially in the medical field. [6].

Pezzolas et al. (2022) focused on the BGMM-OCE method for generating synthetic data. The method is based on a Gaussian mixture and was tested on data from heart patients, generating 30,000 synthetic data sets, and the model outperformed other models. [7].

Luch et al. (2022) focused on the effect of changes in the parameters of the random forest algorithm on the prediction accuracy.. The study relied on classifying individuals' gender based on pain perception data, and the model achieved a certain percentage of accuracy with the default parameters.

However, after tuning parameters such as the number of trees and tree depth, accuracy increased, confirming the importance of tuning parameters in strengthening research hypotheses[8].

2. Related Works

In 2023, Bugarčić et al. examined the impact of trade infrastructure development on exports in Central and Eastern European countries using a panel regression model, which affects the trade balance. The R^2 result was 0.33 within countries and 0.42 between countries.[10] In 2024, Wahab focused on the impact of trade-related infrastructure on the standard modified gravity model, using the Panel IVs technique data with two-step least squares (2SLS) model. The result of R^2 was 0.42 for exports and 0.33 for imports. [11].

In 2023, Novi Aryani et al. examined the impact of economic growth on the trade balance between Indonesia and China for the period 2000-2021. Using an Error Correction Model (ECM), the R^2 result was 0.75[12].

In 2025, Shaker (2025) focused his research on the variation in trade facilitation on the trade volume of the countries participating in the IMEC Corridor, using a gravity model and the ordinary linear regression (OLS) technique. The value of the (Adjusted R^2) coefficient was 0.794. [13].

In 2022 Guan et al. investigated the impact of the Belt and Road Initiative on GDP growth in four ASEAN countries by applying the ordinary linear regression (OLS) method. The adjusted R^2 value in the model was 0.693.[14].

In 2023 Chinn, Meunier, and Stumpner presented their research on global trade forecasting using, three stages: LARS, dimensionality reduction (PCA), and macroeconomic random forests (MRF). The RMSE reduction was between 15% and 33% compared to traditional models, and 26% compared to PCA-OLS. [15]. In 2020, Wochner proposed a new method for forecasting GDP growth by combining economic theories with machine learning techniques. He called it "Dynamic Factor Trees and Forests," explaining the results of the new model, which improved forecast accuracy by 20% compared to the traditional dynamic factor model. [16]. Mervyn et al. (2021) focused in their study on using the Probabilistic Random Forest (PRF) algorithm to enhance the prediction accuracy. The random forest was modified and the results confirmed 17% superiority using bootstrap sampling [17].

3. Problem Statement

Most traditional statistical tools are always based on linear relationships between variables, which makes it difficult to detect complex, nonlinear patterns in large economic data. This study offers machine learning tools, such as random forests, as a more robust and stable solution, as these algorithms are capable of handling nonlinear relationships and complex large data sets. However, these algorithms can face challenges, including class imbalance, limited data, and fixed repetitions, which can reduce model performance. This study used various techniques, including generating synthetic data using Gaussian mixture models (GMMs) and varying algorithm parameters with a different random number generator function. These techniques helped stabilize the model, reduce overfitting, and improve its accuracy in forecasting the trade balance.

Mathematical Formulation:

Let $X \in \mathbb{R}^{n \times d}$ represent the matrix of input features, where nnn is the number of observations and ddd the number of economic variables (e.g., imports, exports, inflation, exchange rate). Let the output label be
$y \in \{0,1\}^n$

where 0 denotes a **deficit** and 1 denotes a **surplus** in the trade balance. The goal is to learn a function

$f: \mathbb{R}^d \rightarrow \{0, 1\}$

that maximizes classification performance metrics (accuracy, precision, recall) under imbalanced and limited data conditions.

To address these issues, this study enhances the RF classifier by:

1. **Generating synthetic data** $X' \sim \text{GMM}(X)$ using Gaussian Mixture Models to expand and balance the dataset.
2. **Tuning RF parameters** (e.g., tree depth, number of estimators) across multiple random initializations to improve robustness.
3. **Training the model over k=30k=30k=30** randomized iterations to ensure stable and generalizable performance.

This hybrid approach is designed to mitigate data irregularities, reduce model variance, and improve

the overall ability of the system to accurately classify trade balance outcomes.

4. Methodology

4.1 Data Collection

A detailed database has been thoughtfully assembled using trusted sources like the World Bank and the International Monetary Fund. It includes 780 monthly records, spanning from 1960 through December 2024, and captures a wide range of economic indicators such as Iraq's GDP, trade balance, exports, imports, population growth, unemployment rates, foreign investment, and government spending. Special emphasis is placed on the trade balance, which is mainly influenced by fluctuations in exports and imports.

4.2 Data Preprocessing

To enhance data diversity and improve the model's accuracy in predicting the Iraqi trade balance, multiple steps were implemented to process the original data and generate synthetic data using a Gaussian Mixture Model (GMM). Initially, the date column was removed from the original dataset due to its indirect amenability to numerical modeling. The variables were then separated into explanatory variables (Features) and a target variable (Trade_Balance). A GMM model with five components (n_components = 5) was trained to learn the probabilistic structure of the original data, allowing for the generation of synthetic data based on the actual distribution of economic indicators. 780 new synthetic samples were generated the total number of records become 1,560, and a standardization technique was also used using a StandardScaler to ensure uniformity of the range of values, improve model performance, and remove missing values.

4.3 Machine Learning classification

To predict the trade balance, a Random Forest model was applied, based on a set of economic indicators, by repeating the process 30 times to ensure the stability of the results and evaluate the performance. In each iteration, the data is randomly divided into a training and test set at a ratio of 80% and 20%. A new random number between 0.1 and 0.9 is also generated to be multiplied by the (base_random_state), which allows for variation in the model. (400) decision trees were also used with other features such as the Gini criterion for node division, not specifying a maximum tree depth, and using log2 to select the number of features in each division. The model was also trained befor and after using Synthetic Data Generation with GMMs. This design contributes to training a robust model capable of capturing changing patterns in the data.

4.4 Model Evaluation Metrics:

The quality of the model was evaluated by applying various criteria, including accuracy, in addition to a full classification report containing precision, recall and The following tables and figers were generated for analysis.

1- Initial Random Forest

Table1 shows the model's performance over 30 runs using accuracy metrics. The overall average accuracy was 0.9665, indicating good and stable model performance. The highest accuracy (0.9872) was achieved in iterations 1, 5, and 11. of these, iteration 11 is the best, combining high accuracy (0.9872), high recall (0.9808), and strong predictive accuracy (0.9808), reflecting good stability in model performance. Figure 4 displays the model's average performance across 30 iterations. The average precision across iterations was 0.9665, demonstrating the model's high ability to correctly classify. The average recall was 0.9381, reflecting the model's strength in retrieving the majority of positive cases, while the average predictive accuracy was 0.9741, indicating that false positives were few The results demonstrated that the model maintained stable and efficient performance, with good balance across all criteria.

2- Enhanced Random Forest

Table 2 displays the performance results of the random forest model on the orginal data and after adjusting its parameters in each run, including the random value and randomness condition, along with changes in parameters such as tree depth or partitioning method. The average accuracy was 0.9694, with a recall of 0.9523 and a predictive accuracy of 0.96982, reflecting consistent performance. The most notable results were achieved by iteration 19, which achieved the highest accuracy (0.9936) with excellent balance across the remaining metrics .The results show that the model stayed effective even when the settings were adjusted, highlighting its ability to adapt and remain stable.Table 3 highlights the model's performance of Random Forest model with adjusting parameters and random value generation after applying Synthetic Data Generation with GMMs.over 30 different runs, each using varied random values and conditions. On average, the model achieved an accuracy of 98.23%, a recall of 98.08%, and a predictive accuracy of 98.85%.

These results reflect the model's strong stability and reliability, with only slight variations between iterations. This consistency confirms the model's ability to accurately classify the data.

5. Results And Discussion

5.1 Model Result

Applying the Random Forest model to forecast the trade balance yielded strong results both before and after doubling the dataset. The model consistently maintained high accuracy across 30 different runs, even as random values and decision trees were varied dynamically. Before data doubling, the model's accuracy reached around 96.94%, demonstrating its solid ability to predict whether the trade balance would show a surplus or a deficit.. Performance remained at the same level after data doubling, with a slight improvement. Accuracy reached 98.23% in several iterations, indicating that the increased sampling was positive and enhanced the model's performance. This contributed to improving the model's stability without compromising its efficiency

Results from 30 Iterations (befor Synthetic Data Generation with GMMs):

- **Accuracy:** 0.9694
- **Recall:** 0.9523
- **Precision:** 0.9682

Figure 5 illustrates the average performance of the model on the original data. The values confirmed the model's strong and stable performance and its ability to diagnose classes, even before using any methods to improve the distribution or balance of the data, confirming the quality of the underlying data structure

Results from 30 Iterations (after Synthetic Data Generation with GMMs):

- **Accuracy:** 0.9823
- **Recall:** 0.9808
- **Precision:** 0.9885

Figure 6 illustrates the model's performance results after applying synthetic data generation using GMMs. The results showed an improvement, which helped enhance the model's performance, increase its ability to generalize when using additional data, and reduce prediction errors.

5.2 Comparison Performance Results:

Table 4 compares the performance of the improved Random Forest model with the K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) algorithms in predicting the trade balance, using both original data and data extended using a GMM. The comparison was conducted using 30 iterations to ensure model stability, in addition to applying a random number generation mechanism to change the random state at each iteration.

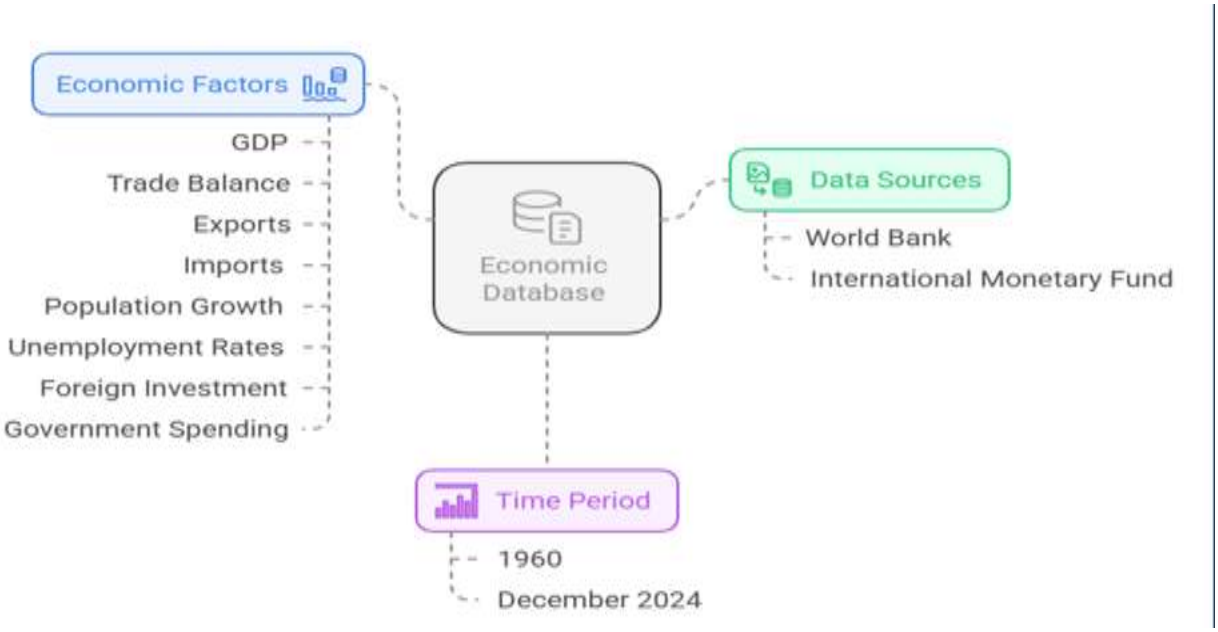


Figure.1 *Iraq Economic data base*

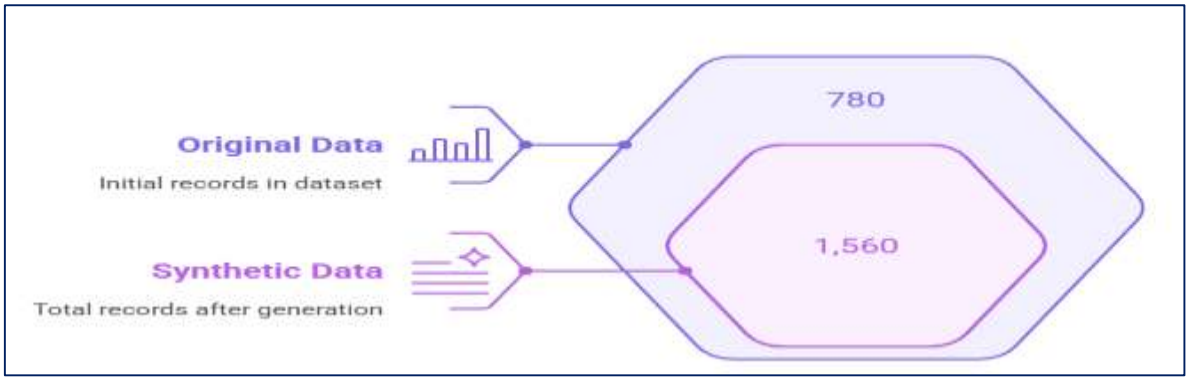


Figure.2 *Data Preprocessing*

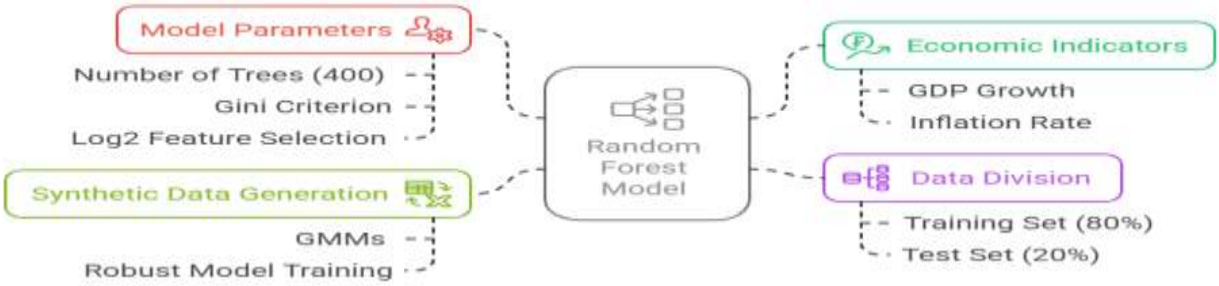


Figure.3 *Random Forest Model*

Table1: *Initial Random Forest metric Result*

Iteration	Accuracy	Recall	Precision
1	0.9872	0.9643	1.0
2	0.9487	0.9118	0.9688
3	0.9744	0.9524	0.9836
4	0.9744	0.9839	0.9531
5	0.9872	0.9697	1.0
6	0.9679	0.9153	1.0
7	0.9679	0.9677	0.9524
8	0.9744	0.9492	0.9825
9	0.9744	0.9444	0.9808
10	0.9487	0.8923	0.9831
11	0.9872	0.9808	0.9808
12	0.9615	0.9355	0.9667
13	0.9615	0.9123	0.9811
14	0.9679	0.9254	1.0
15	0.9744	0.9524	0.9836
16	0.9615	0.9298	0.9636
17	0.9615	0.9153	0.9818
18	0.9679	0.9138	1.0
19	0.9744	0.9492	0.9825
20	0.9744	0.9492	0.9825
21	0.9744	0.9565	0.9851
22	0.9679	0.9344	0.9828
23	0.9808	0.9667	0.9831
24	0.9808	1.0	0.9538
25	0.9808	0.9804	0.9615
26	0.9487	0.8971	0.9839
27	0.9359	0.8621	0.9615
28	0.9423	0.902	0.92
29	0.9295	0.875	0.9245
30	0.9551	0.9552	0.9412
Average	0.9665	0.9381	0.9741

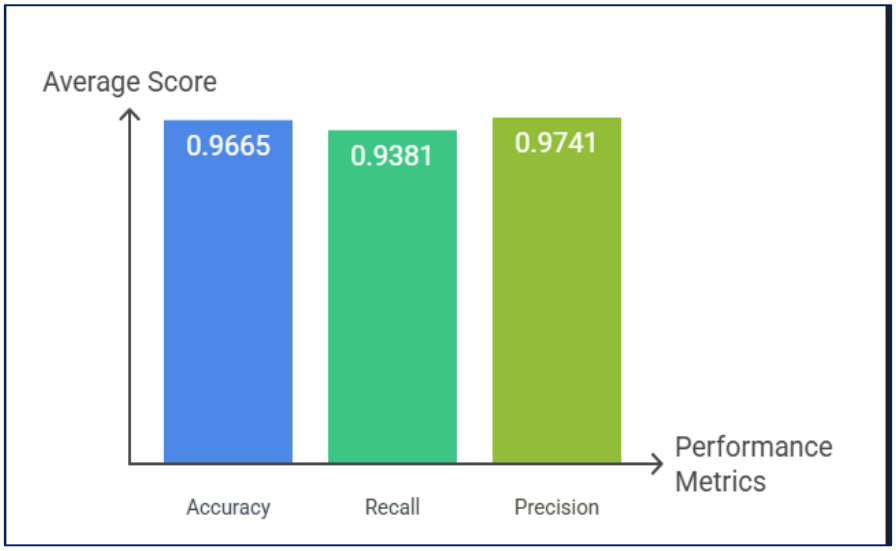


Figure 4. *performance metric Accros Iterration*

Table 2:*Random Forest befor Synthetic Data Generation with GMMs*

Iteration	Random State	Accuracy	Recall	Precision	Random Value
1	19.0	0.9679	1.0	0.9206	0.4623
2	23.0	0.9872	0.9825	0.9825	0.5587
3	6.0	0.9551	0.9167	0.9649	0.1654
4	15.0	0.9872	0.9623	1.0	0.3669
5	28.0	0.9359	0.9219	0.9219	0.6899
6	10.0	0.9744	0.9815	0.9464	0.258
7	4.0	0.9679	0.9344	0.9828	0.1106
8	37.0	0.9615	0.9516	0.9516	0.8819
9	6.0	0.9551	0.9167	0.9649	0.1639
10	11.0	0.9615	0.9219	0.9833	0.2847
11	23.0	0.9872	0.9825	0.9825	0.556
12	21.0	0.9808	0.9434	1.0	0.519
13	23.0	0.9872	0.9825	0.9825	0.565
14	22.0	0.9551	0.9016	0.9821	0.546
15	15.0	0.9872	0.9623	1.0	0.3662
16	22.0	0.9551	0.9016	0.9821	0.5271
17	19.0	0.9679	1.0	0.9206	0.4591
18	10.0	0.9744	0.9815	0.9464	0.2438
19	33.0	0.9936	0.9831	1.0	0.7986
20	4.0	0.9679	0.9344	0.9828	0.11
21	35.0	0.9679	0.9254	1.0	0.8402
22	14.0	0.9487	0.9385	0.9385	0.3456
23	35.0	0.9679	0.9254	1.0	0.8426
24	32.0	0.9551	0.9434	0.9259	0.7754
25	29.0	0.9679	0.9815	0.9298	0.7071
26	7.0	0.9744	0.9844	0.9545	0.1683
27	16.0	0.9679	0.9286	0.9811	0.4019
28	34.0	0.9744	0.9483	0.9821	0.8261
29	7.0	0.9744	0.9844	0.9545	0.1881
30	27.0	0.9744	0.9474	0.9818	0.6607
Average	19.5667	0.9694	0.9523	0.9682	0.4796

***Table** 3:Random Forest aftet Synthetic Data Generation with GMMs*

Iteration	Random State	Accuracy	Recall	Precision	Random Value
1	22.0	0.984	0.9836	0.989	0.532
2	7.0	0.9744	0.9886	0.9667	0.1892
3	20.0	0.9679	0.983	0.9611	0.4965
4	37.0	0.9776	0.9749	0.9898	0.8995
5	10.0	0.9808	0.9728	0.9944	0.2433
6	23.0	0.984	0.9837	0.9891	0.5564
7	19.0	0.9872	0.9944	0.9832	0.4667
8	24.0	0.9872	0.9888	0.9888	0.593
9	32.0	0.9968	0.9944	1.0	0.7806
10	31.0	0.9872	0.9833	0.9944	0.7449
11	32.0	0.9968	0.9944	1.0	0.7699
12	24.0	0.9872	0.9888	0.9888	0.5876
13	12.0	0.9904	0.9896	0.9948	0.2881
14	24.0	0.9872	0.9888	0.9888	0.5949
15	35.0	0.9712	0.9581	0.9877	0.8448
16	9.0	0.9776	0.974	0.9894	0.2224
17	34.0	0.9872	0.9945	0.9838	0.8329
18	11.0	0.9744	0.9738	0.9841	0.2745
19	35.0	0.9712	0.9581	0.9877	0.8353
20	33.0	0.984	0.9766	0.994	0.8093
21	12.0	0.9904	0.9896	0.9948	0.3019
22	22.0	0.984	0.9836	0.989	0.528
23	25.0	0.9776	0.9721	0.9886	0.5988
24	15.0	0.9808	0.9672	1.0	0.3572
25	24.0	0.9872	0.9888	0.9888	0.582
26	35.0	0.9712	0.9581	0.9877	0.8383
27	26.0	0.9808	0.98	0.9899	0.6402
28	24.0	0.9872	0.9888	0.9888	0.581
29	36.0	0.9776	0.9775	0.9831	0.8774
30	9.0	0.9776	0.974	0.9894	0.2371
Average	23.4	0.9823	0.9808	0.9885	0.5701

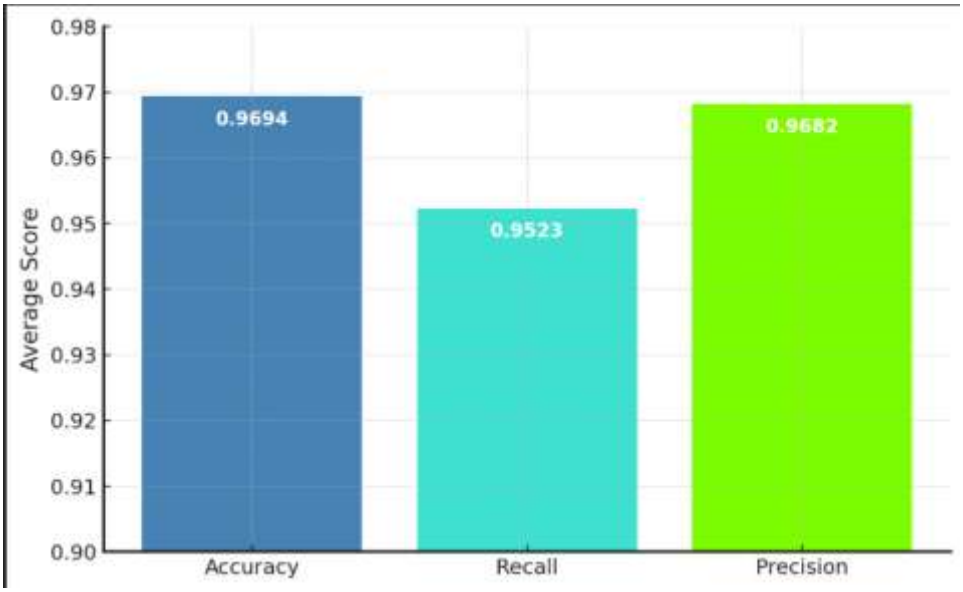


***Figure 5** performance metric Accros Iterration*

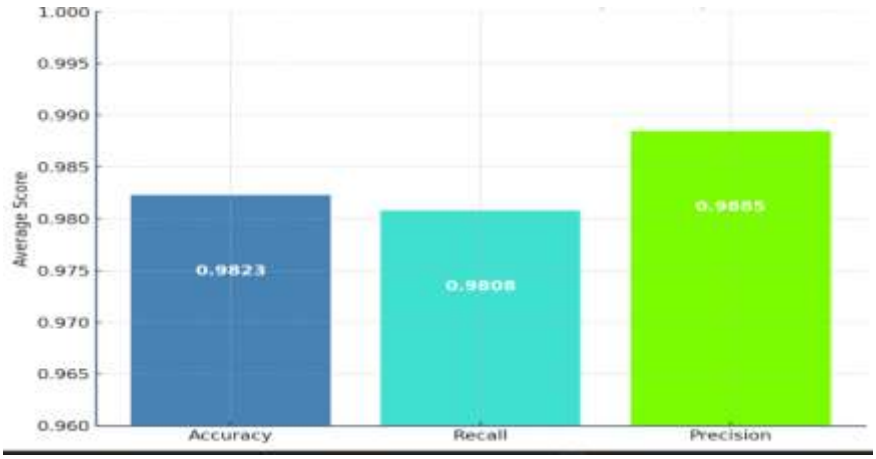


Figure 6. *performance metric Accros Iterration*

Table 4. *Comparison Performance*

	Algorithm Metrices	KNN classifier [18]	SVM classifier [19]	improved Random Forest
before Synthetic Data Generation	Avg Accuracy	0.8972	0.9577	0.9694
	Avg Recall	0.8622	0.9422	0.9523
	Avg Precision	0.8684	0.9465	0.9682
After Synthetic Data Generation	Avg Accuracy	0.9135	0.9691	0.9823
	Avg Recall	0.8246	0.9526	0.9808
	Avg Precision	0.9307	0.9641	0.9885

4. Conclusions

This study confirmed the success of using machine learning techniques with synthetic data generation in accurately predicting the trade balance, especially for limited and variable data. The Random Forest algorithm was used with Gaussian Mixture Models (GMMs) to generate additional synthetic data. The model was implemented through 30 randomized trials using a random number generation function. These techniques helped improve the accuracy and stability of the model, Where the average accuracy changed from 96.94% to 98.23% after generating the synthetic data. Additionally, the accuracy in distinguishing between surplus and deficit improved, increasing from 96.82% to 98.85%. Tuning the model parameters also helped reduce overfitting and make the model less susceptible to bias. These results explains a advanced and flexible model that can be leveraged for data-driven decision-making in trade and economics.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Nuraini, P. (2019). Balance of trade: Theories and practices. *International Journal of Tax Economics and Management*, *2*(1).

[2] Saleh, A. P. D. K. I. (2022). Reality of Foreign Trade Sector in Iraq During the Period 2017-2021. *Tikrit Journal of Administrative and Economic Sciences*, *18*(60 part 1).

[3] Alkooranee, H. T., Assadzadeh, A., & Alsukaini, A. K. M. (2023). An Analysis the Impact of Economic Reform Policies on Liberalizing the Trade Balance in Iraq for the Period 1990-2020. *International Journal of New Political Economy*, *4*(1), 185-211.

[4] Marie, B. S., & Moslem, H. S. (2023). Effect of Public Debt on The Trade Balance in Iraq For The Period (2003-2021). *Journal of Economics and Administrative Sciences*, *29*(136), 40-48.

[5] AL SAEEDI, S. O. A. (2022). The challenges of rail linkage between Iraq and Kuwait and the opportunities for national alternatives (the port of Faw ،the dry canal ،and the link with the Belt and Road project as a model). *Tikrit Journal For Political Science*, *3*(29), 64-95.

[6] Apellániz, P. A., Parras, J., & Zazo, S. (2024, August). An improved tabular data generator with VAE-GMM integration. In *2024 32nd European Signal Processing Conference (EUSIPCO)* (pp. 1886-1890). IEEE.

[7] Pezoulas, V. C., Tachos, N. S., Gkois, G., Olivotto, I., Barlocco, F., & Fotiadis, D. I. (2022). Bayesian inference-based Gaussian mixture models with optimal components estimation towards large-scale synthetic data generation for in silico clinical trials. *IEEE Open Journal of Engineering in Medicine and Biology*, *3*, 108-114.

[8] Lötsch, J., & Mayer, B. (2022). A biomedical case study showing that tuning random forests can fundamentally change the interpretation of supervised data structure exploration aimed at knowledge discovery. *BioMedInformatics*, *2*(4), 544-552..

[10] Bugarčić, F., Mićić, V., & Bošković, G. (2024). THE NEXUS BETWEEN TRADE INFRASTRUCTURE DEVELOPMENT AND EXPORT: THE CASE OF CENTRAL AND EASTERN EUROPEAN COUNTRIES. *TEME*, 907-922.

[11] Wahab, B. A. (2024). Trade-related infrastructure and bilateral trade flows: evidence from Nigeria and its trading partners. *Journal of Economic Structures*, *13*(1), 13.

[12 Ariani, N., & Amaliah, I. (2023). Pengaruh Pertumbuhan Ekonomi, Inflasi, dan Nilai Tukar Terhadap Neraca Perdagangan Indonesia-China. *Jurnal Riset Ilmu Ekonomi Dan Bisnis*, 75-84.

[13] Shaker, S. A. (2025). The impact of trade facilitation inequality on bilateral trade: the case of India–Middle East–Europe economic corridor (IMEC). *Journal of Shipping and Trade*, *10*(1), 7.

[14] Guan, T., Gou, H., & Qi, Y. (2022). THE EFFECT OF THE BELT AND ROAD INITIATIVE ON GDP GROWTH: EVIDENCE FROM MALAYSIA, INDONESIA, THAILAND, AND PHILIPPINES. *European journal of law and political sciences*, (2-3), 37-45..

15] Chinn, M. D., Meunier, B., & Stumpner, S. (2023). *Nowcasting world trade with machine learning: a three-step approach* (No. w31419). National Bureau of Economic Research.

[16] Wochner, D. (2020). Dynamic factor trees and forests–a theory-led machine learning framework for non-linear and state-dependent short-term us gdp growth predictions. *KOF Working Papers*, *472*.

[17 Mervin, L. H., Trapotsi, M. A., Afzal, A. M., Barrett, I. P., Bender, A., & Engkvist, O. (2021). Probabilistic Random Forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. *Journal of Cheminformatics*, *13*, 1-17.

[18] Gazi, M. S., Barua, A., Karim, F., Siddiqui, M. I. H., Das, N., Islam, M. R., ... & Al Montaser, M. A. (2025). Machine Learning-Driven Analysis of Low-Carbon Technology Trade and Its Economic Impact in the USA. *Journal of Ecohumanism*, *4*(1), 4961-4984.

[19] Putri, R. A., Abapihi, B., & Arisona, D. C. (2024). Support Vector Machine: Classification of Trade Balance of Provincies in Indonesia Based on Gross Regional Domestic Product and Large Trade Price Index in 2023. *International Journal of Economics, Management and Accounting*, *1*(2), 221-231.