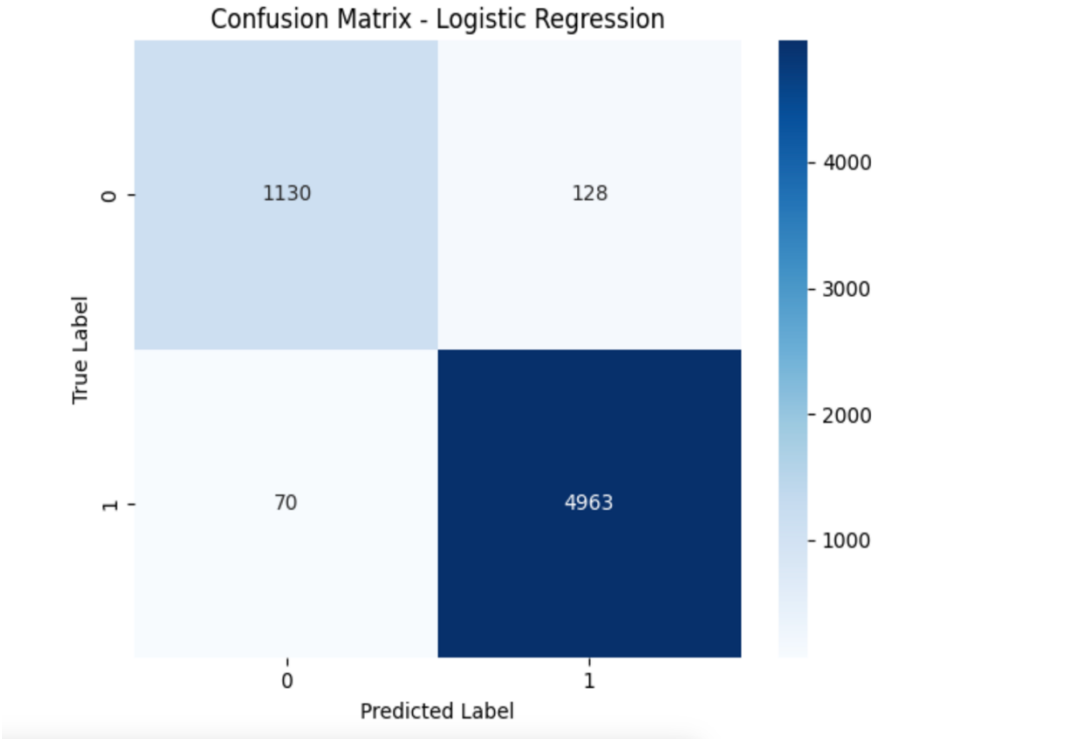
Analysis of the results of Arabic text classification models: Model reports and performance

Introduction:

This report aims to evaluate the performance of several classification models to determine whether Arabic text was written by a human or generated by artificial intelligence. Traditional machine learning models (Logistic Regression, SVM, Random Forest) were compared with a deep learning model (FFNN). The results were presented using standard evaluation metrics (Accuracy, Precision, Recall, F1 score), as well as confusion matrices and training curves.

Logistic Regression:

```
                precision    recall  f1-score   support

...        human       0.94      0.90      0.92      1258
           ai          0.97      0.99      0.98      5033

    accuracy                             0.97      6291
   macro avg       0.96      0.94      0.95      6291
weighted avg       0.97      0.97      0.97      6291

[Saved Figure] /content/Detection-AI-Generated-Arabic-Text/reports/figures/lr_cm.png
```

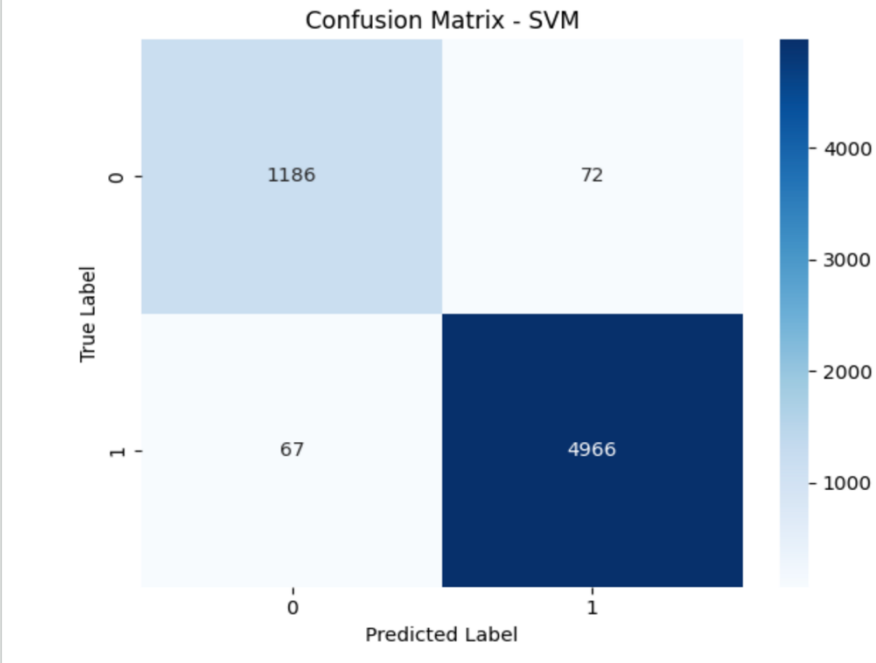


The Logistic Regression model performed exceptionally well in classifying Arabic texts, achieving an overall accuracy of approximately 96.85%.

The confusion matrix shows that the model excels at detecting AI-generated texts (correctly classifying 4,963 AI texts), with a small number of errors, including misclassifying 70 AI texts as human.

Regarding human texts, 1,130 were correctly classified, while 128 were incorrectly classified as AI, indicating a slight bias in classifying some human texts as AI due to a data imbalance (a larger number of AI texts).

SVM:

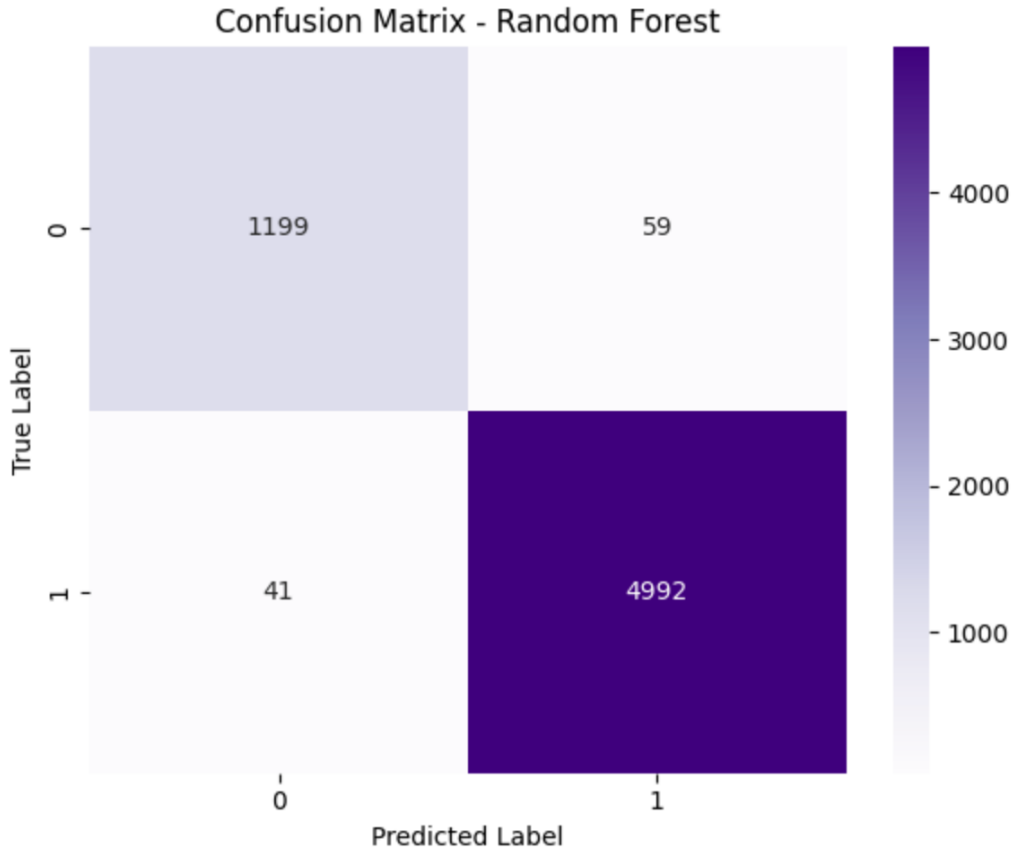



```
Training SVM
SVM Test Accuracy: 0.9779049435701797
Confusion Matrix:
 [[1186   72]
 [  67 4966]]

[SVM] Val Accuracy: 0.9737678855325914
              precision    recall  f1-score   support

       human       0.95      0.94      0.94      1258
          ai       0.99      0.99      0.99      5033

    accuracy                           0.98      6291
   macro avg       0.97      0.96      0.97      6291
weighted avg       0.98      0.98      0.98      6291

[SVM] Test Accuracy: 0.9779049435701797
              precision    recall  f1-score   support

           0       0.95      0.94      0.94      1258
           1       0.99      0.99      0.99      5033

    accuracy                           0.98      6291
   macro avg       0.97      0.96      0.97      6291
weighted avg       0.98      0.98      0.98      6291
```

The SVM model performed exceptionally well in classifying Arabic texts, achieving a test accuracy of approximately 97.79%. The confusion matrix shows that the model correctly classified 4966 AI-generated texts, with only a few errors, misclassifying 67 AI texts as human. It also correctly classified 1186 human texts, while misclassifying 72 human texts as AI. The evaluation metrics demonstrate high Precision/Recall values, particularly for the AI category (≈0.99), reflecting the model's ability to accurately detect generated texts with minimal errors.

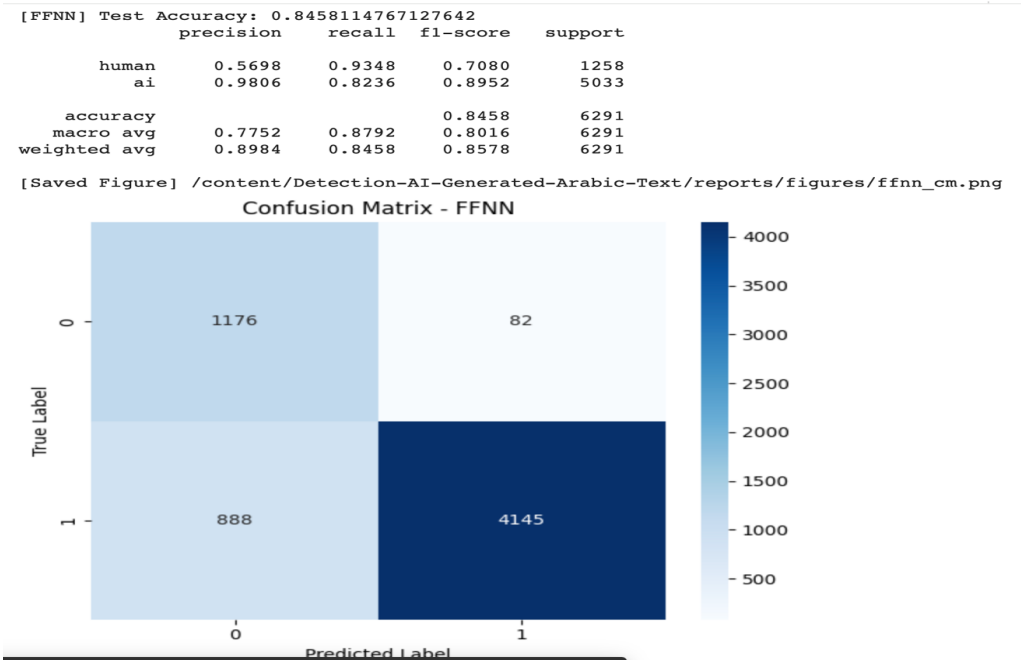
Random Forest

```
Training Random Forest
Confusion Matrix:
 [[1199   59]
 [  41 4992]]

[RF] Val Accuracy: 0.9813990461049285
              precision    recall  f1-score   support

       human       0.96      0.95      0.95      1258
          ai       0.99      0.99      0.99      5032

    accuracy                           0.98      6290
   macro avg       0.97      0.97      0.97      6290
weighted avg       0.98      0.98      0.98      6290

[RF] Test Accuracy: 0.9841042759497695
              precision    recall  f1-score   support

       human       0.97      0.95      0.96      1258
          ai       0.99      0.99      0.99      5033

    accuracy                           0.98      6291
   macro avg       0.98      0.97      0.98      6291
weighted avg       0.98      0.98      0.98      6291
```



The Random Forest model achieved the highest performance among traditional models, with a test accuracy of approximately 98.41%. The confusion matrix shows that the model correctly classified 4992 AI-generated texts with a very limited number of errors (41 AI texts were misclassified as Human). It also correctly classified 1199 human texts, while 59 human texts were incorrectly classified as AI. The evaluation metrics confirm high Precision and Recall for the AI category (≈0.99), as well as strong performance for the Human category (F1 ≈0.96), indicating the model's ability to minimize errors in both categories.

FFNN:



The FFNN model achieved a test accuracy of approximately 84.58%, which is lower than the performance of traditional models in this experiment. The confusion matrix shows that the model correctly classified 4145 AI-generated texts, but misclassified a significant number of AI texts as human (888 cases), indicating a lower recall for the AI category compared to other models. Conversely, the model achieved a high recall for the human category (≈0.93) but a low precision for the human category (≈0.57), meaning that a considerable portion of the texts it predicted as human were actually AI. Overall, the results suggest that FFNN needs further refinement (such as improving the text representation, threshold, or network structure) to reach the performance level of traditional models.