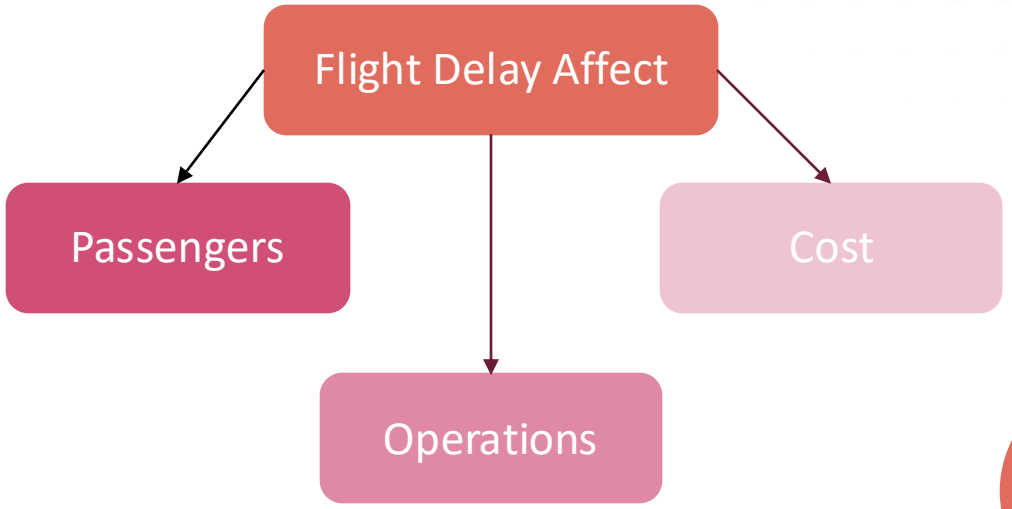
UNIVERSITI TEKNOLOGI MALAYSIA

FLIGHT DELAY PREDICTION MODEL USING MACHINE LEARNING

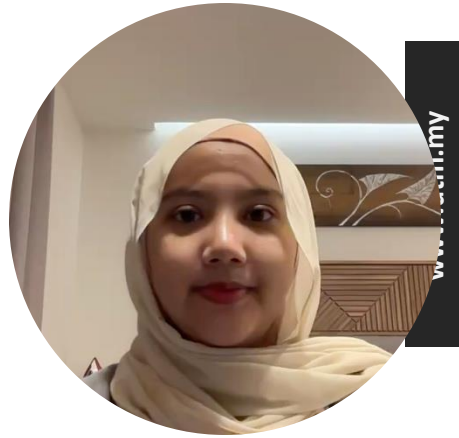
Name: Siti Nur Elisya Aqmar Binti Mohamad Kamal
Matric No: MCS241056
Faculty of Computing, UTM
Date: June 2025

INTRODUCTION

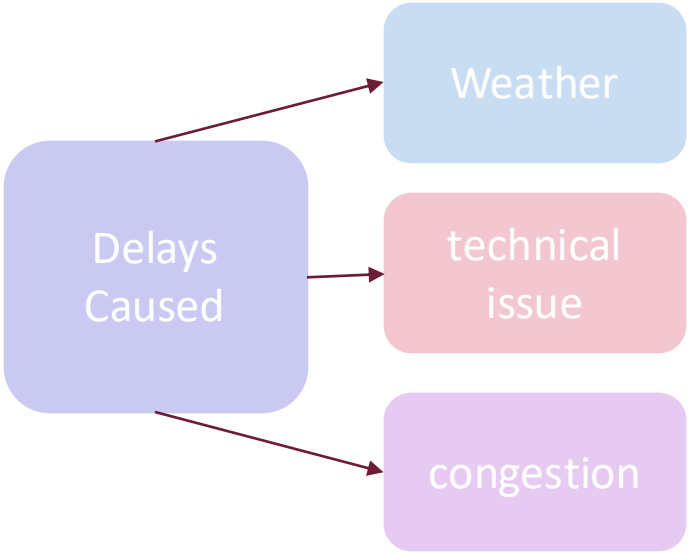


Accurate delay prediction = improves airline efficiency

Machine Learning (ML) is ideal for handling complex aviation data.



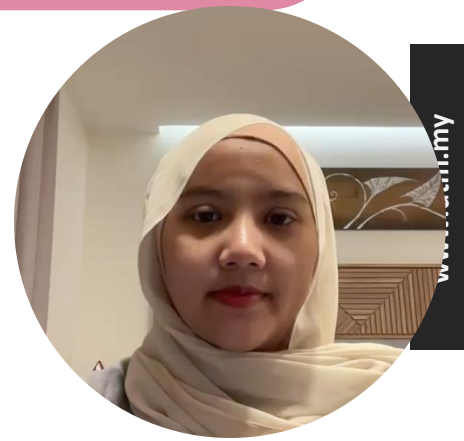
PROBLEM BACKGROUND



Traditional methods lack of accuracy and adaptability

PROBLEM STATEMENT

- Current systems are reactive, not predictive.
- Need for data-driven solutions using ML.
- Goal: Forecast delays proactively for better decision-making.



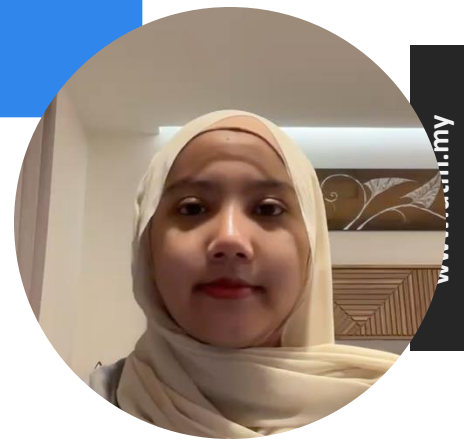
RESEARCH GOAL & OBJECTIVE

To build and compare machine learning models to predict flight delays

Collect and preprocess flight/weather data

Apply and compare Random Forest, XGBoost, ATT-BI-LSTM.

Identify best model via evaluation metrics.



SCOPE OF RESEARCH

Focus on historical structured data (Nov 2019–Dec 2020).

Excludes real-time or unscheduled flight data.

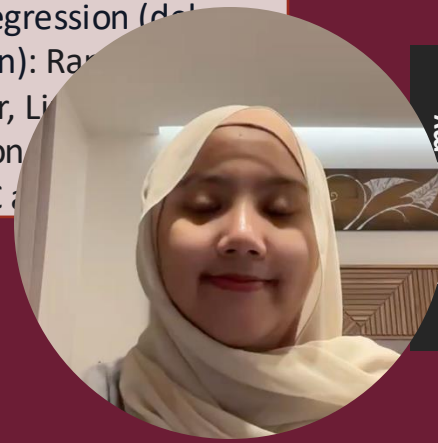
Supervised ML techniques only.

REPORT STRUCTURE

- Chapter 1: Introduction
- Chapter 2: Literature Review
- Chapter 3: Research Methodology
- Chapter 4: Results & Discussion
- Chapter 5: Conclusion & Recommendation

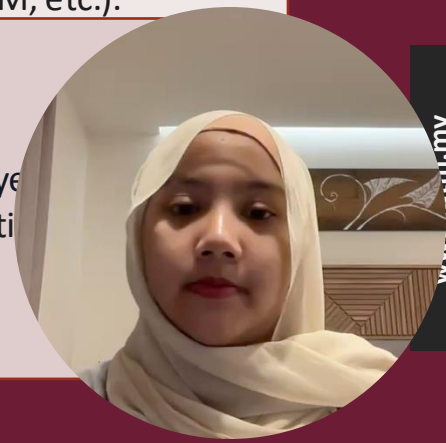
LITERATURE REVIEW

Author	Research Title	Year	Research Focus	Methods
Yazdi, M., et al.	Flight Delay Prediction Based on Deep Learning and Levenberg-Marquardt Algorithm	2020	U.S. flight data (5 years of historical flight schedules) from theBureau of Transportation Statistics. Includes 27 attributes (e.g., flight dates, delays, weather, airports, taxi times).	Deep Learning and Levenberg-Marquardt Algorithm
Zhe Zheng, Wenbin Wei, Minghua Hu	A Comparative Analysis of Delay Propagation on Departure and Arrival Flights for a Chinese Case Study	2021	Flight-level records (2016) from the Air Traffic Management Bureau of CAAC, covering 1,469,909 flights between top 30 Chinese airports. METAR weather reports(temperature, wind speed, visibility, convective weather) for these airports.	Econometric models: Ordinary Least Squares (OLS) regression with clustered
Zhan Shu	Analysis of Flight Delay and Cancellation Prediction Based on Machine Learning Model	2021	Kaggle dataset(2018 U.S. domestic flights): Includes 27 attributes (e.g., flight dates, delays, weather, airline IDs, airports, taxi times). Focus on 2018 data (subset of 2009–2018 dataset).	Random Forest Classifier, Logistic Regression, Naive Bayes. Regression (de… prediction): Ra… Regressor, Li… Regression… ROC-AUC a…

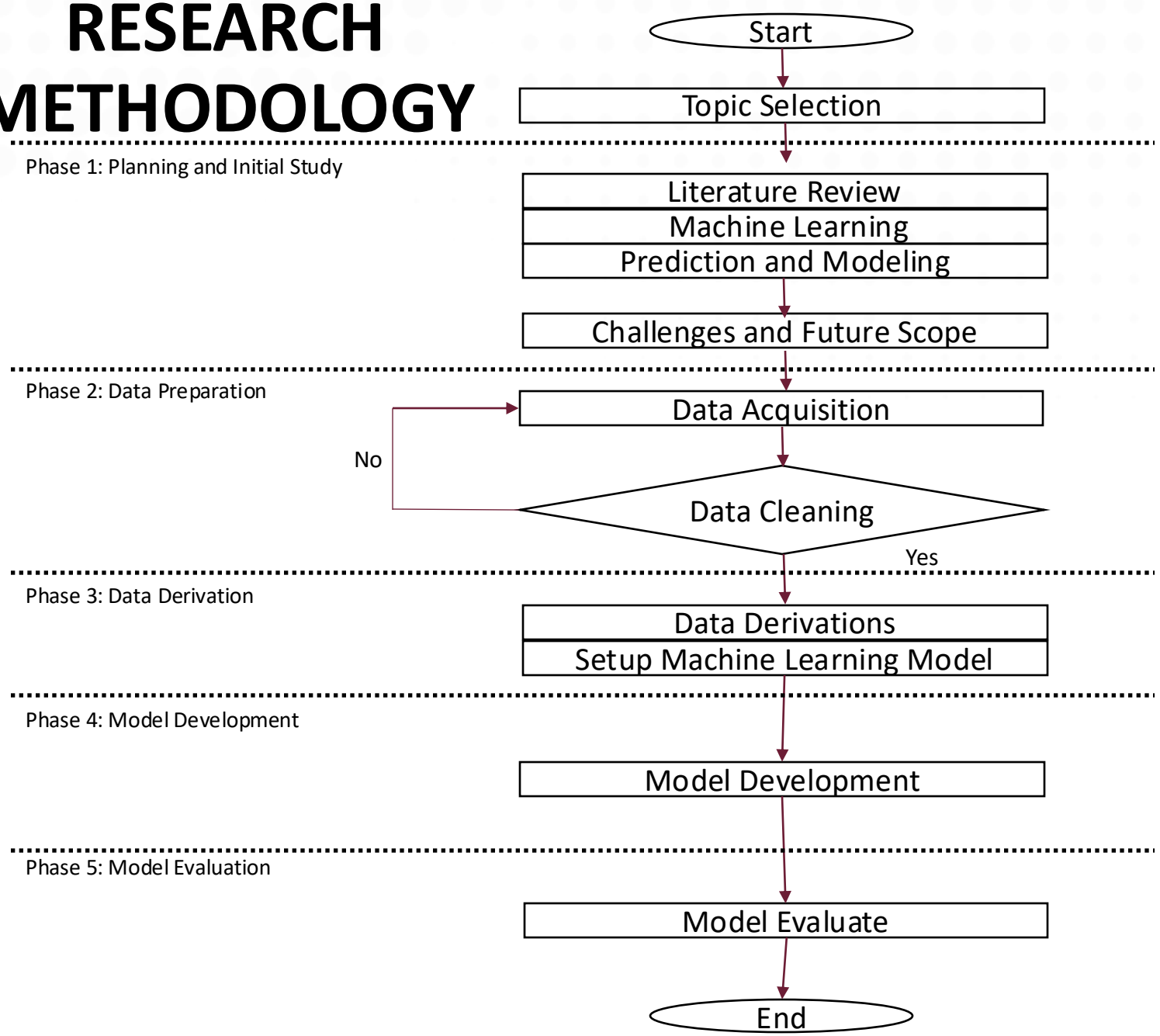


LITERATURE REVIEW

Author	Research Title	Year	Research Focus	Methods
Gui, G., et al. (School of Electronic and Information Engineering, Beihang University, China)	Flight Delay Prediction Based on Aviation Big Data and Machine Learning	2019	ADS-B messages integrated with weather, flight schedules, and airport information. Timeframe: December 2018–May 2019. Size: 5,761 flight records (3,368 no-delay, 2,393 delayed). Coversall routes/airports within the ADS-B platform.	LSTM-based architectures: Standard LSTM, LSTM with fully connected layers, LSTM with dropout layers. Random Forest classifier.
Mamdouh, M., Ezzat, M., Hefny, H., et al. (Department of Computer Science, Helwan University, Egypt)	Improving Flight Delays Prediction by Developing Attention-Based Bidirectional LSTM Network	2024	US flight data (2019–2020) integrated with weather features(wind, temperature, precipitation.	Primary model:Attention-based Bidirectional LSTM (ATT-BI-LSTM). Comparative models: GRU, RNN, LSTM, ATT-LSTM, Deep LSTM, and hybrid models (PSO+LSTM, ACO+LSTM, etc.).
Ahmad Adib Baihaqi Shukri, Syarifah Adilah Mohamed Yusoff, Saiful Nizam Warris, Mohd Saifulnizam Abu Bakar, Rozita Kadar	Machine Learning Approach of Predicting Airline Flight Delay Using Naive Bayes Algorithm	2024	2020 flight dataset with 28,821 samples and 23 attributes (e.g., departure/arrival delays, weather, flight schedules).	Naive Baye classificati



RESEARCH METHODOLOGY



This research framework includes the following steps:

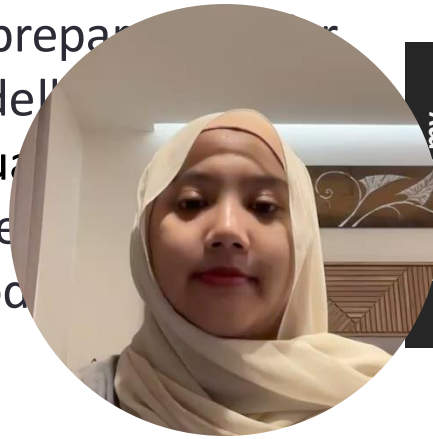
1.Problem Definition and Literature Review

2. Data Collection: Gather flight and weather data specific to Malaysia.

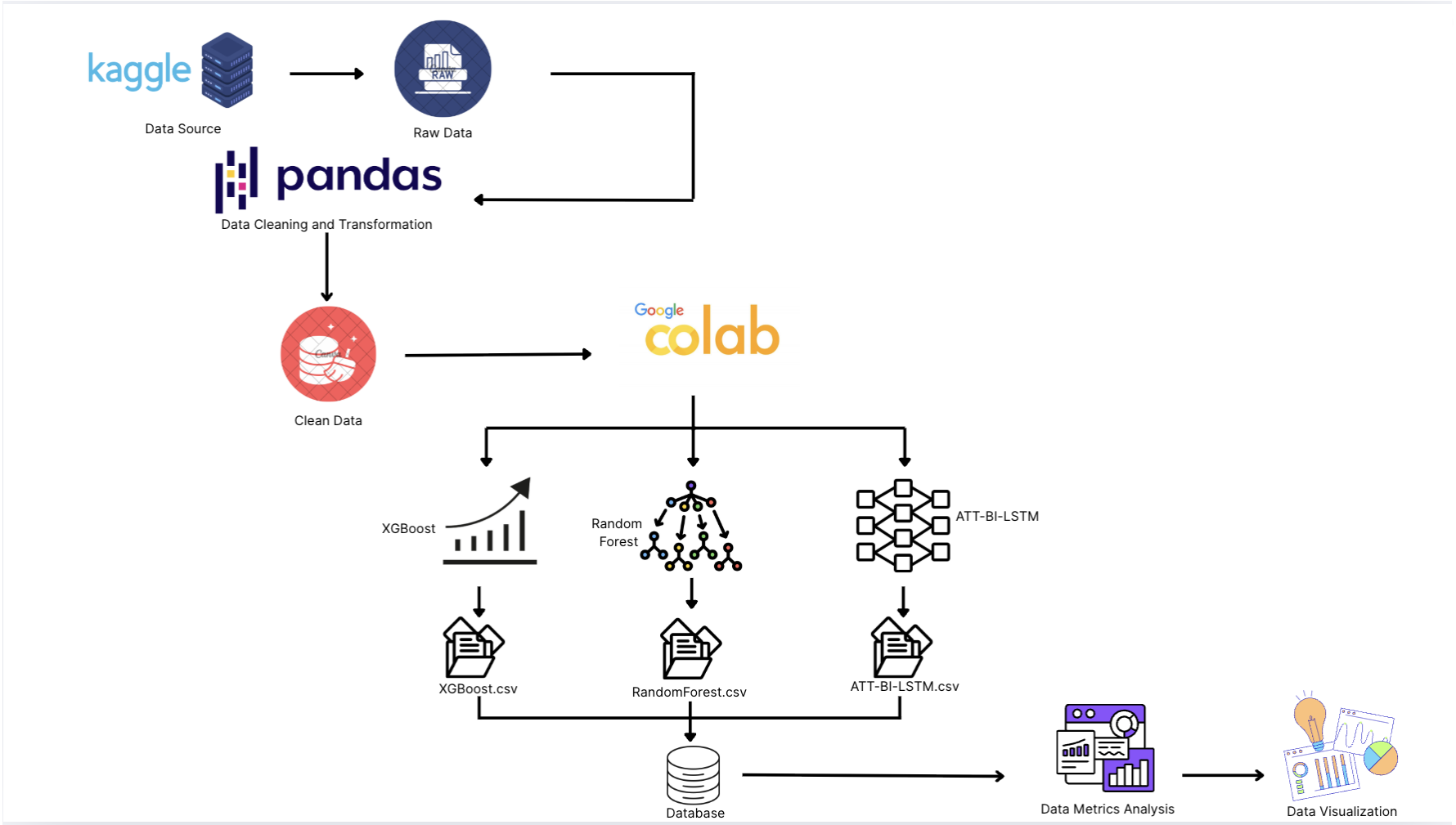
3. Data Pre-processing: Clean, balance, and prepare data for modelling

4. Model Development: Clean, balance, and prepa… r… model…

5. Model Evalua… performance… mod…



DATA PREPROCESSING



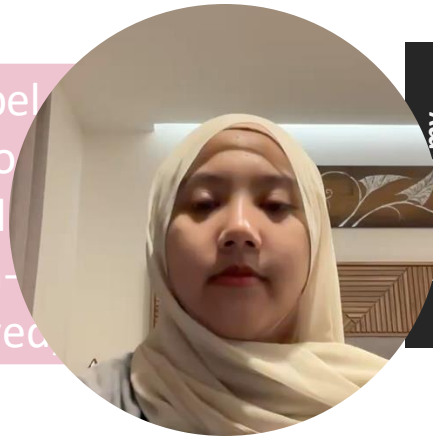
OVERVIEW OF DATASET

Source: M1_final.csv from Kaggle

Total records: 28821, with 23 features

Key attributes: departure time, airline, route, weather, delay label

Delay label distributio
imbalanced d
(e.g., 70% on-
30% delayed



DATA EXPLORATION

M1_final

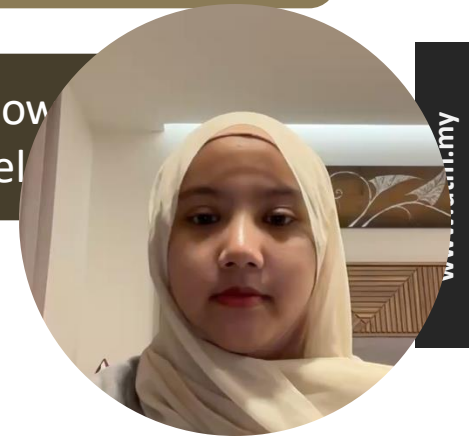
MONTH	DAY_OF_MONTH	DAY_OF_WEEK	OP_UNIQUE_CARRIER	TAIL_NUM	DEST	DEP_DELAY	CRS_ELAPSED_TIME	DISTANCE	CRS_DEP_M	DEP_TIME_M	CRS_ARR_M	Temperature	Dew Point	Humidity	Wind	Wind Speed	Wind Gust	Pressure	Condition	sch_dep	sch_arr	TAXI_OUT
11	1	5	B6	N828JB	CHS	-1	124	636	324	323	448	48	34	58	W	25	38	29.86	Fair / Windy	9	17	14
11	1	5	B6	N992JB	LAX	-7	371	2475	340	333	531	48	34	58	W	25	38	29.86	Fair / Windy	9	17	15
11	1	5	B6	N959JB	FLL	40	181	1069	301	341	482	48	34	58	W	25	38	29.86	Fair / Windy	9	17	22
11	1	5	B6	N999JQ	MCO	-2	168	944	345	343	513	48	34	58	W	25	38	29.86	Fair / Windy	9	17	12
11	1	5	DL	N880DN	ATL	-4	139	760	360	356	499	46	32	58	W	24	35	29.91	Fair / Windy	9	17	13
11	1	5	AA	N983NN	ORD	-1	161	740	359	358	460	46	32	58	W	24	35	29.91	Fair / Windy	9	17	21
11	1	5	AA	N107NN	LAX	-1	373	2475	360	359	553	46	32	58	W	24	35	29.91	Fair / Windy	9	17	26
11	1	5	B6	N274JB	BUF	-5	80	301	365	360	445	46	32	58	W	24	35	29.91	Fair / Windy	17	21	11
11	1	5	B6	N663JB	LGB	0	368	2465	365	365	553	46	32	58	W	24	35	29.91	Fair / Windy	17	21	25
11	1	5	B6	N283JB	FLL	3	184	1069	370	373	554	46	32	58	W	24	35	29.91	Fair / Windy	17	21	29
11	1	5	B6	N962JT	LAS	-5	343	2248	381	376	544	46	32	58	W	24	35	29.91	Fair / Windy	17	21	26
11	1	5	AA	N901AN	DCA	-5	95	213	384	379	479	46	32	58	W	24	35	29.91	Fair / Windy	17	21	30
11	1	5	AA	N157UW	PHX	-4	336	2153	390	386	546	46	32	58	W	24	35	29.91	Fair / Windy	17	21	24
11	1	5	B6	N967JT	SFO	-3	388	2586	410	407	618	46	32	58	W	24	35	29.91	Fair / Windy	17	21	17
11	1	5	B6	N998JE	SJU	108	222	1598	301	409	523	46	32	58	W	24	35	29.91	Fair / Windy	17	21	16
11	1	5	DL	N703TW	SFO	-6	391	2586	420	414	631	47	33	59	W	24	29	30	Fair / Windy	17	21	16
11	1	5	DL	N192DN	SLC	-5	321	1990	419	414	620	47	33	59	W	24	29	30	Fair / Windy	17	21	25
11	1	5	DL	N362NB	BOS	-5	82	187	420	415	502	47	33	59	W	24	29	30	Fair / Windy	17	21	15
11	1	5	AA	N115NN	SFO	-3	386	2586	420	417	626	47	33	59	W	24	29	30	Fair / Windy	17	21	18
11	1	5	B6	N638JB	SAV	-2	143	718	419	417	562	47	33	59	W	24	29	30	Fair / Windy	17	21	14
11	1	5	B6	N292JB	SYR	-3	73	209	420	417	493	47	33	59	W	24	29	30	Fair / Windy	17	21	21
11	1	5	B6	N806JB	ATL	0	154	760	418	418	572	47	33	59	W	24	29	30	Fair / Windy	17	21	16
11	1	5	DL	N119DU	MSP	-2	187	1029	420	418	547	47	33	59	W	24	29	30	Fair / Windy	17	21	34
11	1	5	B6	N968JT	LAX	-7	379	2475	425	418	624	47	33	59	W	24	29	30	Fair / Windy	17	21	15
11	1	5	DL	N179DN	LAX	-1	365	2475	420	419	605	47	33	59	W	24	29	30	Fair / Windy	17	21	23
11	1	5	DL	N722TW	SEA	-2	373	2422	425	423	618	47	33	59	W	24	29	30	Fair / Windy	30	26	24
11	1	5	AA	N585UW	MIA	4	184	1089	420	424	604	47	33	59	W	24	29	30	Fair / Windy	30	26	24
11	1	5	AS	N526AS	SEA	-6	380	2422	430	424	630	47	33	59	W	24	29	30	Fair / Windy	30	26	19
11	1	5	AS	N557AS	PDX	-3	375	2454	445	442	640	47	33	59	W	24	29	30	Fair / Windy	30	26	20
11	1	5	B6	N584JB	TPA	-2	180	1005	447	445	627	47	33	59	W	24	29	30	Fair / Windy	30	26	13
11	1	5	B6	N273JB	BTV	6	75	266	455	461	530	47	33	59	W	24	29	30	Fair / Windy	30	26	11
11	1	5	B6	N324JB	ORD	-4	163	740	467	463	570	47	33	59	W	24	29	30	Fair / Windy	30	26	13
11	1	5	AA	N116AN	LAX	-6	376	2475	480	474	676	50	33	52	W	21	30	30	Fair / Windy	30	26	20
11	1	5	AA	N844NN	ORD	-3	166	740	479	476	585	50	33	52	W	21	30	30	Fair / Windy	30	26	19
11	1	5	DL	N774DE	IAH	-3	253	1417	480	477	673	50	33	52	W	21	30	30	Fair / Windy	30	26	23
11	1	5	DL	N920DU	DEN	-2	271	1626	480	478	631	50	33	52	W	21	30	30	Fair / Windy	30	26	20
11	1	5	B6	N661JB	RSW	29	193	1074	449	478	642	50	33	52	W	21	30	30	Fair / Windy	30	26	10
11	1	5	B6	N983JT	LAX	-1	376	2475	480	479	676	50	33	52	W	21	30	30	Fair / Windy	30	26	12
11	1	5	MQ	N853AE	ORF	-6	88	290	485	479	573	50	33	52	W	21	30	30	Fair / Windy	30	26	25
11	1	5	9E	N923XJ	JAX	-5	156	828	485	480	641	50	33	52	W	21	30	30	Fair / Windy	46	19	24
11	1	5	B6	N535JB	MCO	0	176	944	480	480	656	50	33	52	W	21	30	30	Fair / Windy	46	19	10
11	1	5	DL	N322US	MIA	-5	181	1089	485	480	666	50	33	52	W	21	30	30	Fair / Windy	46	19	36
11	1	5	DL	N328NB	MSY	-4	189	1182	485	481	614	50	33	52	W	21	30	30	Fair / Windy	46	19	11
11	1	5	AA	N181UW	CLT	-4	136	541	485	481	621	50	33	52	W	21	30	30	Fair / Windy	46	19	21
11	1	5	DL	N338DN	MCO	-3	170	944	485	482	655	50	33	52	W	21	30	30	Fair / Windy	46	19	15
11	1	5	9E	N933XJ	BNA	-2	159	765	485	483	584	50	33	52	W	21	30	30	Fair / Windy	46	19	24
11	1	5	9E	N349PQ	RDU	-2	111	427	485	483	596	50	33	52	W	21	30	30	Fair / Windy	46	19	12
11	1	5	DL	N712TW	SFO	-7	400	2586	490	483	710	50	33	52	W	21	30	30	Fair / Windy	46	19	28
11	1	5	B6	N789JB	FLL	4	187	1069	480	484	667	50	33	52	W	21	30	30	Fair / Windy	46	19	25
11	1	5	DL	N850DN	SJU	-1	234	1598	485	484	719	50	33	52	W	21	30	30	Fair / Windy	46	19	29
11	1	5	DL	N706TW	SAN	0	359	2446	485	485	664	50	33	52	W	21	30	30	Fair / Windy	46	19	25
11	1	5	DL	N900DU	SJC	-6	388	2569	495	489	703	50	33	52	W	21	30	30	Fair / Windy	46	19	26

Peak delays observed during morning and late evening

Higher delays on specific weekdays

Certain airlines/routes show consistently higher delay rates

Weather variables show
with delay likel



DATA PREPROCESSING INSIGHTS

```python
import pandas as pd
from sklearn.preprocessing import LabelEncoder

def clean_flight_data(filepath):
    # Load dataset
    df = pd.read_csv(filepath)
    print("✅ Loaded data with shape:", df.shape)
```

pandas: Used for loading, manipulating, and cleaning structured data (dataframes).
LabelEncoder: Converts categorical string values into numerical values for machine learning.

Defines a reusable function called clean_flight_data that takes the path to a CSV file as input.

- Loads the CSV into a DataFrame.
- Prints the number of rows and columns (shape) to confirm successful loading.
- Lists columns expected to contain numerical data.

```python
    numeric_cols = ['Temperature', 'Dew Point', 'Humidity', 'Wind Speed', 'Wind Gust', 'Pressure', 'TAXI_OUT']
```

```python
    for col in numeric_cols:
        if col in df.columns:
            df[col] = pd.to_numeric(df[col], errors='coerce')
            df[col].fillna(df[col].mean(), inplace=True)
```

- Loops through these columns:
  - Converts values to numbers (coerce turns invalid values into NaNs).
  - Replaces any missing (NaN) values with the column's mean value.

```python
    # Fill missing values for 'Condition' (categorical)
    if 'Condition' in df.columns:
        df['Condition'].fillna('Unknown', inplace=True)
```

Fills missing values in the 'Condition' column with the label 'Unknown'.

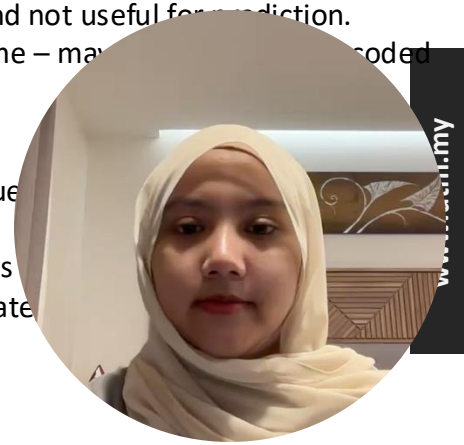
```python
    # Optional: Drop high-cardinality or irrelevant columns
    drop_cols = ['TAIL_NUM', 'sch_dep', 'sch_arr']
    df.drop(columns=[col for col in drop_cols if col in df.columns], inplace=True)
```

Defines a list of columns to remove:
- TAIL_NUM: Aircraft tail number – often unique and not useful for prediction.
- sch_dep, sch_arr: Scheduled departure/arrival time – may [illegible] coded versions elsewhere.

Drops them if present.

```python
    # Encode categorical columns
    cat_cols = ['OP_UNIQUE_CARRIER', 'DEST', 'Condition']
    for col in cat_cols:
        if col in df.columns:
            le = LabelEncoder()
            df[col] = le.fit_transform(df[col].astype(str))
```

- Specifies columns with categorical string value[illegible]
- For each listed column:
  - Converts the data to string (in case it's
  - Uses LabelEncoder to convert each cate[illegible]



```
    # Create delay classification label
    if 'DEP_DELAY' in df.columns:
        df['is_delayed'] = df['DEP_DELAY'].apply(lambda x: 1 if x > 15 else 0)

    # Drop remaining rows with missing values (if any)
    df.dropna(inplace=True)
```

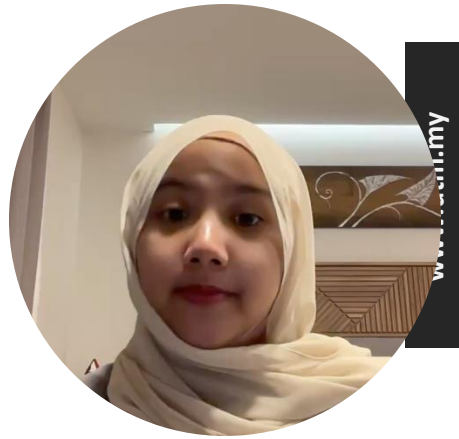
Adds a new column is_delayed:
•If the departure delay is more than 15 minutes, it marks the flight as delayed (1), otherwise not delayed (0).
•This column becomes the target label for classification.
•Removes any rows that still contain missing values.

```
    # Final report
    print("✅ Final cleaned shape:", df.shape)
    print("🧹 Missing values per column:\n", df.isnull().sum())
    return df
```

•Displays the final dataset shape.
•Shows the count of any remaining missing values by column (should be zero).

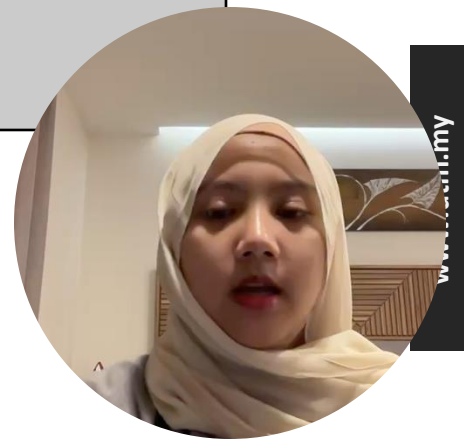
```
# Usage
cleaned_df = clean_flight_data('M1_final.csv')

# Optional: Save to new CSV
cleaned_df.to_csv('cleaned_flight_data.csv', index=False)
print("💾 Cleaned dataset saved as 'cleaned_flight_data.csv'")
```

Calls the function and applies it to the 'M1_final.csv' file.
Saves the cleaned DataFrame to a new CSV file.
Confirms the file has been saved



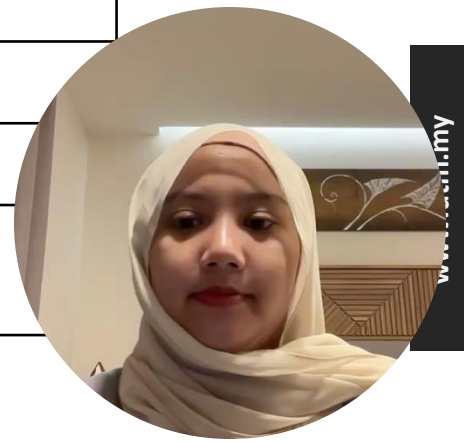
MODEL-SPECIFIC OBSERVATION

Machine Learning	Observation
ATT-BI-LSTM	Strong recall and sequence learning
Random Forest	Good accuracy, interpretable, less effective on time data
XGBoost	Balanced performance, robust to noise, great AUC



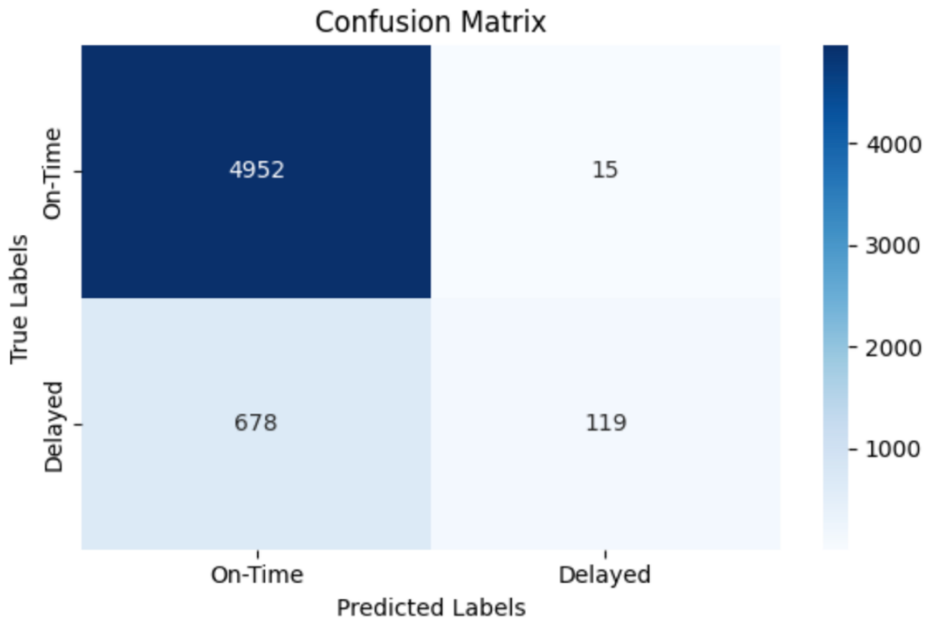
COMPARATIVE PERFORMANCE TABLE

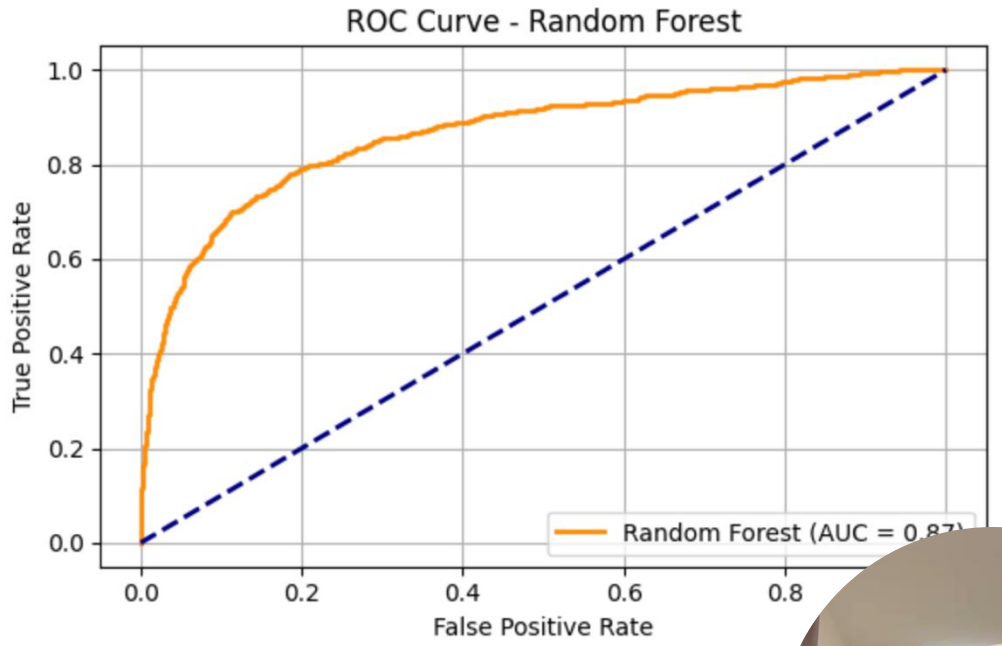
Metric	Random Forest	XGBoost	ATT-BI-LSTM
Accuracy	0.88	1.00	0.90
Precision (Class 1)	0.89	1.00	0.98
Recall (Class 1)	0.15	1.00	0.26
F1-Score (Class 1)	0.26	1.00	0.41
Precision (Class 0)	0.88	1.00	0.89
Recall (Class 0)	1.00	1.00	1.00
F1-Score (Class 0)	0.93	1.00	0.94
Macro Avg F1-Score	0.60	1.00	0.68
Weighted Avg F1	0.84	1.00	0.87
Support (Total Samples)	5764	5764	5764

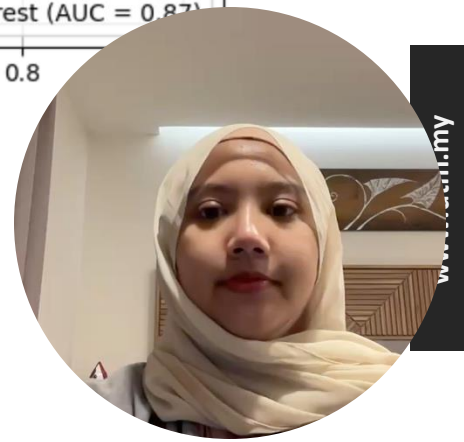


CONFUSION MATRIX & ROC

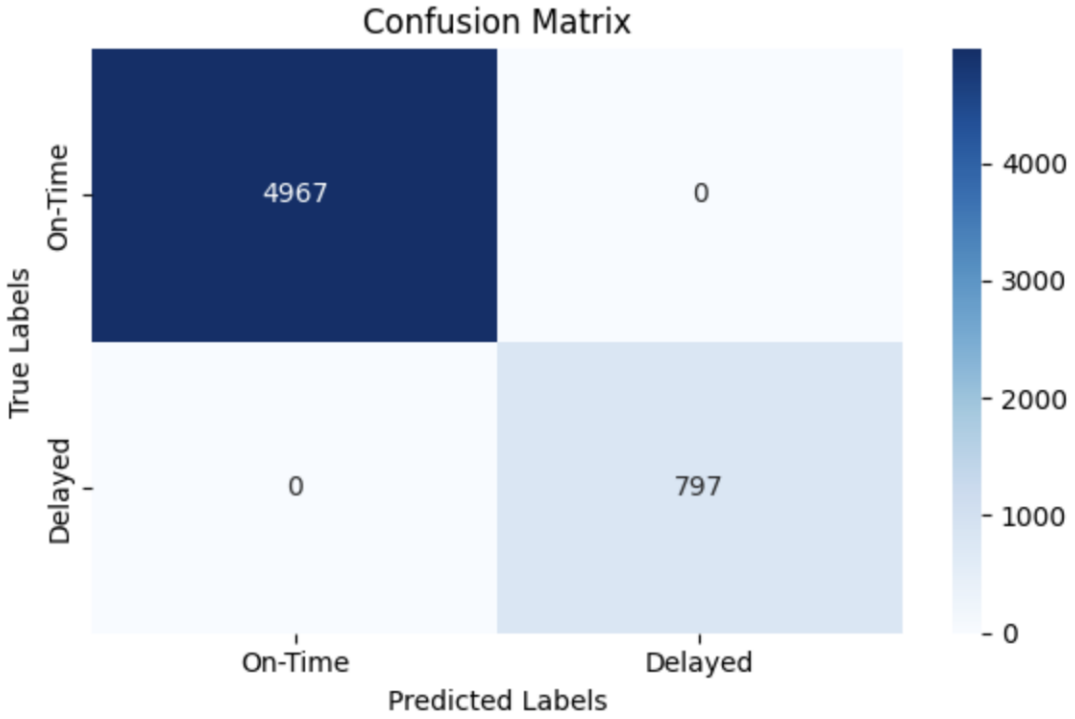
RANDOM FOREST

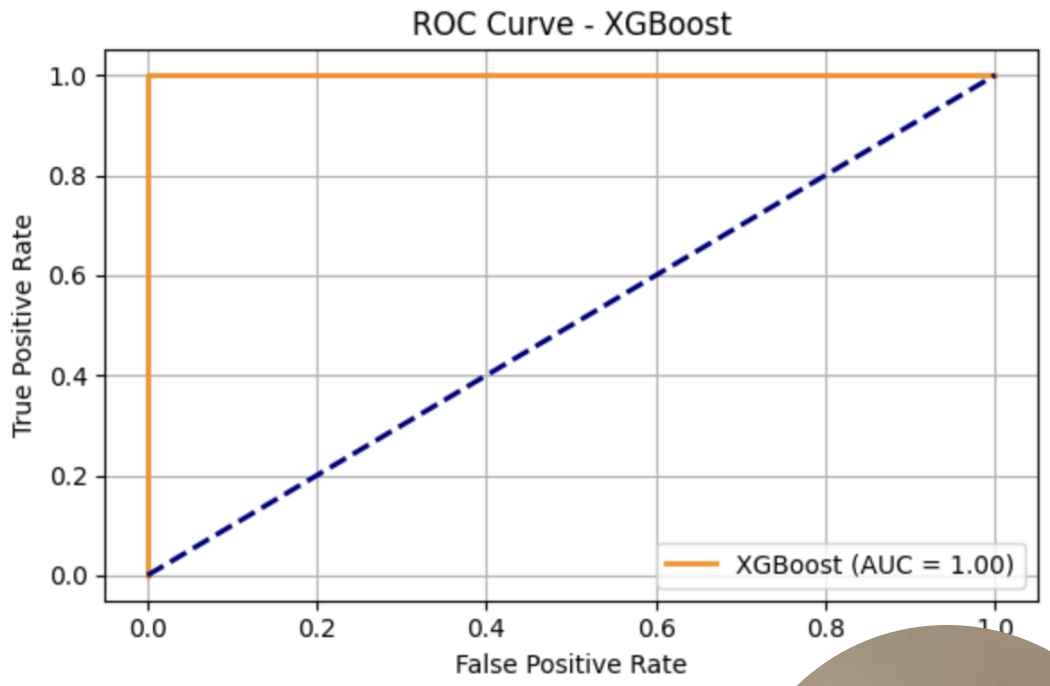


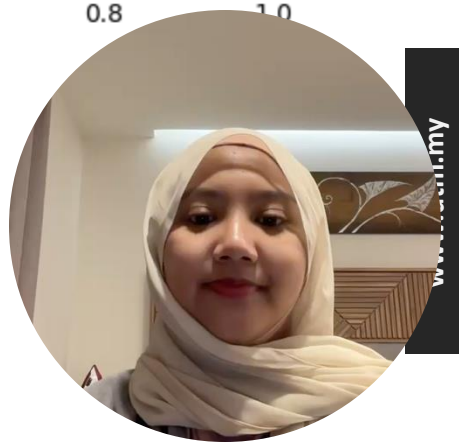




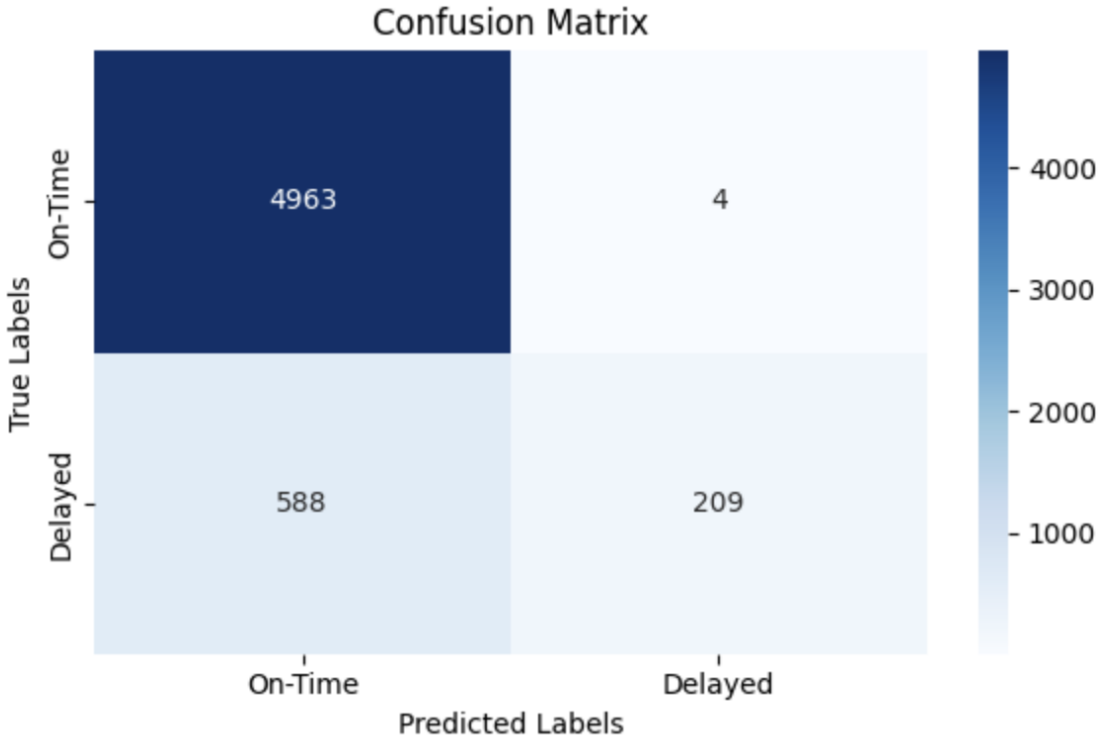
XGBOOST

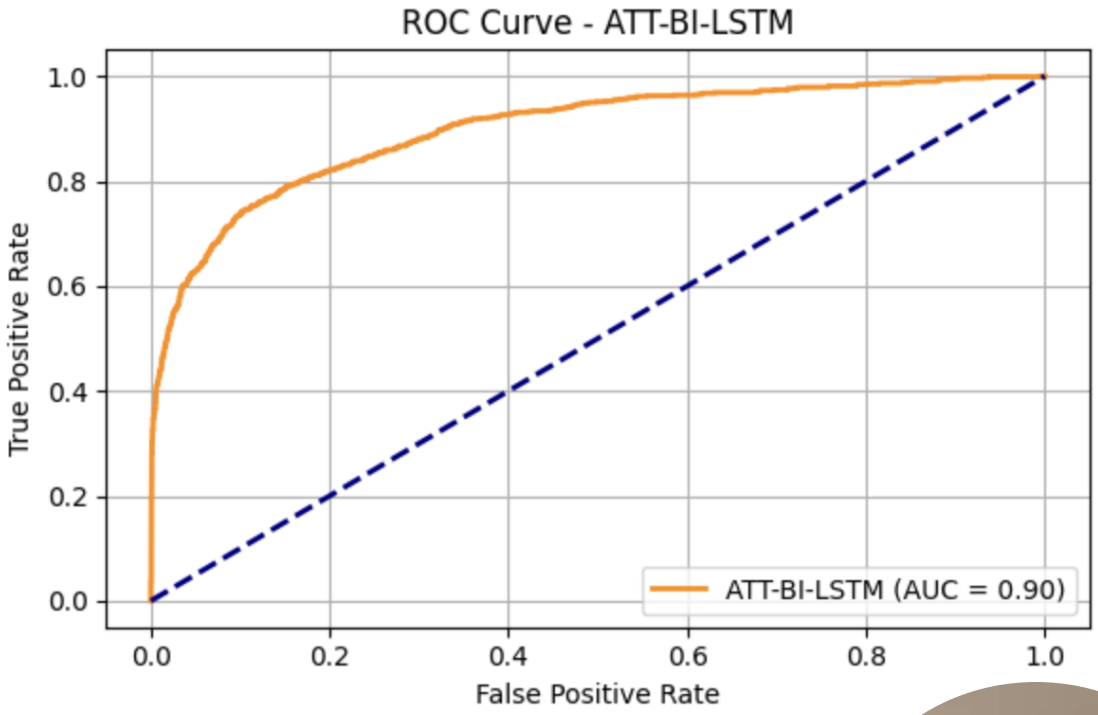

Confusion Matrix
4967
0
0
797
On-Time
Delayed
True Labels
On-Time
Delayed
Predicted Labels
4000
3000
2000
1000
0

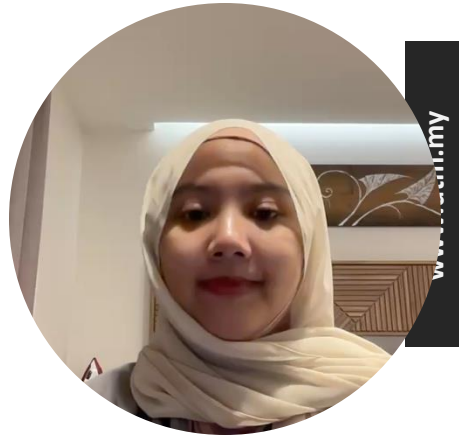

ROC Curve - XGBoost
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
XGBoost (AUC = 1.00)



ATT-BI-LSTM

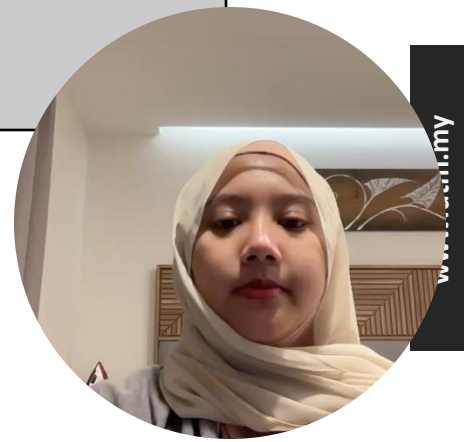

Confusion Matrix
On-Time
Delayed
True Labels
4963
4
588
209
On-Time
Delayed
Predicted Labels
4000
3000
2000
1000


ROC Curve - ATT-BI-LSTM
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ATT-BI-LSTM (AUC = 0.90)



KEY TAKEAWAYS

Machine Learning	Results
ATT-BI-LSTM	best for detecting delays in sequence-based data
Random Forest	suitable when model transparency is important
XGBoost	offers strong overall balance of performance and efficiency



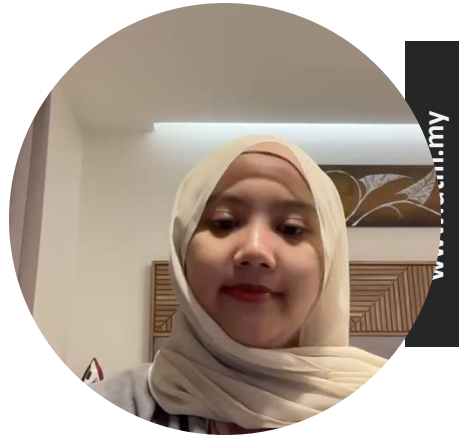
CONCLUSION

ML improves delay prediction, especially XGBoost and ATT-BI-LSTM.

Data preprocessing is critical

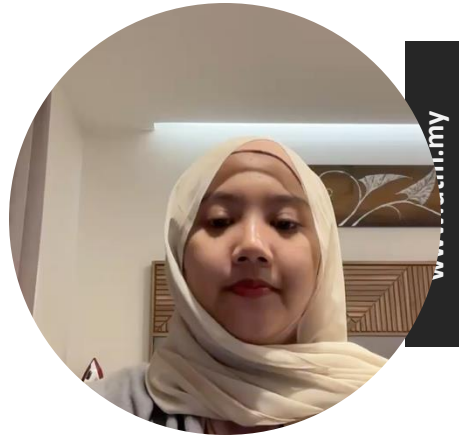
Time and weather are key influencing factors.

ATT-BI-LSTM better at identifying delayed flights



RECOMMENDATIONS & FUTURE WORK

- Add real-time data and multiple airport inputs.
- Try transformer models like BERT, TFT.
- Use Explainable AI (XAI) to improve trust and understanding
- Integrate live system for dynamic predictions



THANK YOU

univteknologimalaysia utm.my utmofficial

Video Presentation Link:
https://youtu.be/KhbUoRJMH0Y