Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)

CSE 498: Literature Review Records

Student's Id and Name	**Name:** Md.Rashed Ul Islam and **ID:** 19202103202
Capstone Project Title	Flight Fare Prediction using Machine Learning
Supervisor Name & Designation	**Name:** Md. Saifur Rahman & **Designation:** Assistant Professor & Chairman, Department of CSE, BUBT
Course Teacher's Name & Designation	**Name:** Khan Md. Hasib & **Designation:** Assistant Professor, Department of CSE, BUBT

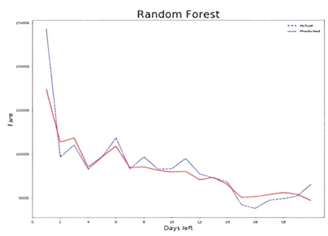
<table>
<tr><th>Aspects</th><th>Paper # 1 (Title)</th></tr>
<tr><td>Title / Question
(What is problem statement?)</td><td>FLIGHT FARE PREDICTION USING MACHINE LEARNING ALGORITHM</td></tr>
<tr><td>Objectives / Goal (What is looking for?)</td><td>The goal of the research paper is to create a machine learning algorithm for flight fare prediction and identify the best times and dates to buy airline tickets at the most affordable prices by utilizing a variety of factors and AI models.</td></tr>
<tr><td>Methodology / Theory (How to find the solution?)</td><td>The research paper utilizes machine learning algorithms, specifically Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), for flight fare prediction
• SVM is used for regression analysis, and the performance depends on the selection of kernel features.
• KNN is used as a non-parametric method for regression analysis, where the result is the mean of the k nearest neighbors.
• The paper also emphasizes the cleaning and preparation of the collected dataset, including the removal of duplicates and invalid values.</td></tr>
<tr><td>Software Tools (What program/software is used for design, coding and simulation?)</td><td>Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.</td></tr>
<tr><td>Test / Experiment How to test and characterize the design/prototype?</td><td>For the experimental work, the datasets were divided into the ratio of 80used to train classification algorithms, and the remaining 20% used for testing purposes. Accuracy, specificity, sensitivity, and area under the curve were evaluated for the seven classifiers.</td></tr>
<tr><td>Simulation/Test Data (What parameters are determined?)</td><td>Datasets was collected from Kaggle.
<table>
<tr><th></th><th>Total_Stops</th><th>Price</th><th>Journey_day</th><th>Journey_month</th><th>Dep_hour</th><th>Dep_min</th><th>Arrival_hour</th><th>Arrival_min</th><th>Duration_hours</th><th>Duration_mins</th><th>Airline_Air India</th><th>Airline_Go</th></tr>
<tr><td>0</td><td>0</td><td>3897</td><td>24</td><td>3</td><td>22</td><td>20</td><td>1</td><td>10</td><td>2</td><td>50</td><td>0</td><td></td></tr>
<tr><td>1</td><td>2</td><td>7662</td><td>1</td><td>5</td><td>5</td><td>50</td><td>13</td><td>15</td><td>7</td><td>25</td><td>1</td><td></td></tr>
<tr><td>2</td><td>2</td><td>13882</td><td>9</td><td>6</td><td>9</td><td>25</td><td>4</td><td>25</td><td>19</td><td>0</td><td>0</td><td></td></tr>
<tr><td>3</td><td>1</td><td>6218</td><td>12</td><td>5</td><td>18</td><td>5</td><td>23</td><td>30</td><td>5</td><td>25</td><td>0</td><td></td></tr>
<tr><td>4</td><td>1</td><td>13302</td><td>1</td><td>3</td><td>16</td><td>50</td><td>21</td><td>35</td><td>4</td><td>45</td><td>0</td><td></td></tr>
<tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td></td></tr>
<tr><td>10678</td><td>0</td><td>4107</td><td>9</td><td>4</td><td>19</td><td>55</td><td>22</td><td>25</td><td>2</td><td>30</td><td>0</td><td></td></tr>
<tr><td>10679</td><td>0</td><td>4145</td><td>27</td><td>4</td><td>20</td><td>45</td><td>23</td><td>20</td><td>2</td><td>35</td><td>1</td><td></td></tr>
<tr><td>10680</td><td>0</td><td>7229</td><td>27</td><td>4</td><td>8</td><td>20</td><td>11</td><td>20</td><td>3</td><td>0</td><td>0</td><td></td></tr>
<tr><td>10681</td><td>0</td><td>12648</td><td>1</td><td>3</td><td>11</td><td>30</td><td>14</td><td>10</td><td>2</td><td>40</td><td>0</td><td></td></tr>
<tr><td>10682</td><td>2</td><td>11753</td><td>9</td><td>5</td><td>10</td><td>55</td><td>19</td><td>15</td><td>8</td><td>20</td><td>1</td><td></td></tr>
</table>
10682 rows × 30 columns</td></tr>
<tr><td>Result / Conclusion (What was the final result?)</td><td>Observed a maximum accuracy of 94.62%.


Figure 3: Random Forest</td></tr>
<tr><td>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</td><td>Limited data availability, variability in ticket prices, overfitting and underfitting challenges, incorporating real-time data, exploring ensemble methods, feature selection techniques, and considering external factors for improved prediction accuracy.</td></tr>
<tr><td>Terminology (List the common basic words frequently used in this research field)</td><td>Machine learning algorithms, SVM, decision trees, KNN, MSE, R-squared, bagging trees, random forests, data analysis, feature selection.</td></tr>
</table>

<table>
<tr><th>Aspects</th><th>Paper # 2 (Title)</th></tr>
<tr><td>Title / Question
(What is problem statement?)</td><td>A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques</td></tr>
<tr><td>Objectives / Goal (What is looking for?)</td><td>The objective of the research study is to investigate machine learning algorithms for predicting flight prices and to find common pricing practices among various airline businesses. The study examines machine learning (ML), deep learning (DL), and quantum machine learning (QML), three different AI model domains. The goal is to offer the end user the most inexpensive ticket price while taking both airline- and destination-based evaluations into consideration.</td></tr>
<tr><td>Methodology / Theory (How to find the solution?)</td><td>The methodology involves extracting a set of effective features from flight data of Aegean, Turkish, Austrian, and Lufthansa Airlines for popular international destinations. Three different domains of AI models are considered: Machine Learning (ML),Deep Learning (DL), and Quantum Machine Learning (QML). A total of 16 model architectures, including eight state-of-the-art ML models, six CNN models in DL, and two QML models, are used to resolve the airfare price prediction problem. The evaluation is conducted from both a destination-based and an airline-based perspective.</td></tr>
<tr><td>Software Tools (What program/software is used for design, coding and simulation?)</td><td>Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.</td></tr>
<tr><td>Test / Experiment How to test and characterize the design/prototype?</td><td>For the experimental work, the datasets were divided into the ratio of 80used to train classification algorithms, and the remaining 20% used for testing purposes. Accuracy, specificity, sensitivity, and area under the curve were evaluated for the seven classifiers.</td></tr>
<tr><td>Simulation/Test Data (What parameters are determined?)</td><td>The datasets are collected from Aegean, Turkish, Austrian, and Lufthansa Airlines for six popular international destinations
<table>
<tr><th>Airline</th><th colspan="7">Destination</th></tr>
<tr><th></th><th>AMS</th><th>ARN</th><th>BRU</th><th>CDG</th><th>LIS</th><th>VIE</th><th>All</th></tr>
<tr><td>Aegean</td><td>17754</td><td>17756</td><td>15427</td><td>18374</td><td>16598</td><td>9727</td><td>95435</td></tr>
<tr><td>Lufthansa</td><td>4219</td><td>5328</td><td>4787</td><td>3851</td><td>5628</td><td>6092</td><td>29900</td></tr>
<tr><td>Turkish</td><td>1515</td><td>1129</td><td>1124</td><td>1765</td><td>1288</td><td>1583</td><td>8391</td></tr>
<tr><td>Austrian</td><td>850</td><td>373</td><td>285</td><td>607</td><td>607</td><td>1010</td><td>3191</td></tr>
</table></td></tr>
<tr><td>Result / Conclusion (What was the final result?)</td><td>Experimental results show that at least three models from each domain (ML, DL, and QML) achieve accuracies between 89% and 99% in predicting airfare prices for different international destinations and airline companies. The study compares the performance of ML models with QML models and finds that QML models, specifically Quantum Machine Learning Perceptron (QMLP), outperform classical ML models like MLP and SVM in terms of enhanced performance.</td></tr>
<tr><td>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</td><td>The research paper does not mention any methodological obstacles or challenges during the study. Future work could include studying the same airline companies and destinations from different airports to examine if the information can be efficiently extracted. Additionally, the research proposes studying the problem as a classification problem through customer segmentation based on the flight features set.</td></tr>
</table>

<table>
<tr><th>Aspects</th><th>Paper # 3 (Title)</th></tr>
<tr><td>Title / Question
(What is problem statement?)</td><td>Implementation of Flight Fare Prediction System Using Machine Learning</td></tr>
<tr><td>Objectives / Goal (What is looking for?)</td><td>The objective of the project is to provide effective flight prices for customers and make the user interface user-friendly. It also aims to gain more exposure to machine learning techniques and improve existing skills. It focuses on using data mining techniques, such as clustering and linear regression, to predict flight fares.</td></tr>
<tr><td>Methodology / Theory (How to find the solution?)</td><td>The paper utilizes data mining techniques, such as clustering and linear regression, to predict flight fares. The authors collect and analyze data on flight prices. Clustering is used to group flights into different categories based on their pricing. Linear regression is employed to develop a predictive model for flight fares, taking into account various factors and parameters from the dataset . The project also focuses on creating a user-friendly interface to enhance the user experience.</td></tr>
<tr><td>Software Tools (What program/software is used for design, coding and simulation?)</td><td>Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.</td></tr>
<tr><td>Test / Experiment How to test and characterize the design/prototype?</td><td>The performance of the flight fare prediction system may have been evaluated using metrics such as Mean Absolute Error (MAE) and R-square. For the experimental work, the datasets were divided into the ratio of 80% and 20%. 80% were used to train and the remaining 20% used for testing purposes.</td></tr>
<tr><td>Simulation/Test Data (What parameters are determined?)</td><td>Datasets was collected from - UCI repository.
<table>
<tr><th></th><th>Airline</th><th>Date_of_Journey</th><th>Source</th><th>Destination</th><th>Route</th><th>Dep_Time</th><th>Arrival_Time</th><th>Duration</th><th>Total_Stops</th><th>Additional_Info</th><th>Price</th></tr>
<tr><td>0</td><td>IndiGo</td><td>24/03/2019</td><td>Banglore</td><td>New Delhi</td><td>BLR → DEL</td><td>22:20</td><td>01:10 22 Mar</td><td>2h 50m</td><td>non-stop</td><td>No info</td><td>3897</td></tr>
<tr><td>1</td><td>Air India</td><td>1/05/2019</td><td>Kolkata</td><td>Banglore</td><td>CCU → IXR → BBI → BLR</td><td>05:50</td><td>13:15</td><td>7h 25m</td><td>2 stops</td><td>No info</td><td>7662</td></tr>
<tr><td>2</td><td>Jet Airways</td><td>9/06/2019</td><td>Delhi</td><td>Cochin</td><td>DEL → LKO → BOM → COK</td><td>09:25</td><td>04:25 10 Jun</td><td>19h</td><td>2 stops</td><td>No info</td><td>13882</td></tr>
<tr><td>3</td><td>IndiGo</td><td>12/05/2019</td><td>Kolkata</td><td>Banglore</td><td>CCU → NAG → BLR</td><td>18:05</td><td>23:30</td><td>5h 25m</td><td>1 stop</td><td>No info</td><td>6218</td></tr>
<tr><td>4</td><td>IndiGo</td><td>01/03/2019</td><td>Banglore</td><td>New Delhi</td><td>BLR → NAG → DEL</td><td>16:50</td><td>21:35</td><td>4h 45m</td><td>1 stop</td><td>No info</td><td>13302</td></tr>
</table></td></tr>
<tr><td>Result / Conclusion (What was the final result?)</td><td>Prediction accuracy of 0.869 with the adjusted R squared performance metrics, with the lowest error rate of 0.92% using the XGBoost algorithm.Another accuracy rates of different machine learning techniques and mentions achieving an accuracy rate of 81.8% with the Trend Based Model method.</td></tr>
<tr><td>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</td><td>Limited availability and quality of flight fare data may pose challenges in accurately predicting flight fares. The scalability and generalizability of the flight fare prediction system across different routes and airlines in the Indian Domestic Airline market may require additional research and development. Mitigating defects such as handling outliers or missing data, can be a challenge that requires continuous improvement and refinement.</td></tr>
<tr><td>Terminology (List the common basic words frequently used in this research field)</td><td>Flight fare prediction ,Machine learning techniques, Data mining, Regression analysis, Clustering, Support Vector Machine (SVM), Random Forest, Gradient boosting, R-squared</td></tr>
</table>

<table>
<tr><th>Aspects</th><th>Paper # 4 (Title)</th></tr>
<tr><td>Title / Question
(What is problem statement?)</td><td>Civil airline fare prediction with a multi- attribute dual-stage attention mechanism</td></tr>
<tr><td>Objectives / Goal (What is looking for?)</td><td>The goal of the research paper is to compare the performance of the Multi-Attribute Dual-stage Attention (MADA) model with other airfare prediction models based on the Auto-Regressive Integrated Moving Average (ARIMA), random forest, or deep learning models. The results of extensive experiments showed that the MADA model outperformed the other models in terms of MSE, RMSE, and MAE indicators.</td></tr>
<tr><td>Methodology / Theory (How to find the solution?)</td><td>The execution process of the MADA model consists of Algorithm 1 (Multi-attribute Data Processing) and Algorithm 2 (Multi-attribute Dual-stage Attention Mechanism model). Algorithm 1 preprocesses the attributes based on their types, normalizes numerical data, and encodes non-numerical data. Algorithm 2 uses the preprocessed data as input and updates the learning parameters through back-propagation to improve the model's generalization capability.</td></tr>
<tr><td>Software Tools (What program/software is used for design, coding and simulation?)</td><td>Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.</td></tr>
<tr><td>Test / Experiment How to test and characterize the design/prototype?</td><td>The data set used in the experiments was a two-year anonymous airfare record from a real airline, containing more than 1.7 million data pieces. The training set was constructed using a portion of the data set, and it contained more than 1.7 million data pieces.</td></tr>
<tr><td>Simulation/Test Data (What parameters are determined?)</td><td>Datasets was collected from - Kaggle.
<table>
<tr><th>Feature attribute</th><th>Description</th></tr>
<tr><td>Airln_cd</td><td>Airline</td></tr>
<tr><td>AirCrft_Typ</td><td>Aircraft type</td></tr>
<tr><td>Dpt_AirPt_Cd</td><td>Departure airfield</td></tr>
<tr><td>Arrv_Airpt_Cd</td><td>Arrival airport</td></tr>
<tr><td>Air_route</td><td>Route</td></tr>
<tr><td>Flt_nbr</td><td>Flight no.</td></tr>
<tr><td>Flt_Schd_Dpt_Tm</td><td>Flight take-off time</td></tr>
<tr><td>Weekday</td><td>Weekly attributes</td></tr>
<tr><td>Holiday</td><td>Holiday</td></tr>
<tr><td>Pax_Qty_y</td><td>Number of flights</td></tr>
<tr><td>Fare</td><td>Air ticket price</td></tr>
</table></td></tr>
<tr><td>Result / Conclusion (What was the final result?)</td><td>The research paper concludes that the proposed Multi-Attribute Dual-stage Attention (MADA) model outperforms other airfare prediction models based on indicators such as Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The MADA model demonstrated better stability and more preferable performance under different time windows compared to other models like CNN-LSTM and Seq2Seq. The prediction results of the MADA model on different routes were at least 2.3% better than the other compared models.</td></tr>
<tr><td>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</td><td>The paper identifies the challenges faced by traditional airfare prediction systems, including the nonlinear interrelationship of multiple factors and the inability to handle the impact of different time steps. The future work includes the exploration of more accurate prediction methods to optimize the current imperfections in airfare prediction.</td></tr>
</table>

<table>
<tr><th>Aspects</th><th>Paper # 5 (Title)</th></tr>
<tr><td>Title / Question
(What is problem statement?)</td><td>Deep-Learning-Powered GRU Model for Flight Ticket Fare Forecasting</td></tr>
<tr><td>Objectives / Goal (What is looking for?)</td><td>The research aims to identify salient and influential features from a large amount of data for financial prediction. The Gated Recurrent Unit (GRU) model is used in this research to capture correlations between flight ticket fares and relevant features. The model utilizes various features, including prices of the same itinerary, recent itineraries, and itineraries for the same day of the week or month. The experimental results demonstrate the superior performance of the proposed model based on ensemble learning</td></tr>
<tr><td>Methodology / Theory (How to find the solution?)</td><td>The GRU model leverages its unique architecture to capture temporal dependencies in-flight data, resulting in improved predictive performance. It addresses the limitations of traditional machine learning techniques that rely heavily on statistical variables in their models.The proposed model outperforms classic machine learning models, as well as MLP and LSTM, in terms of assessment indicators such as mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R2).</td></tr>
<tr><td>Software Tools (What program/software is used for design, coding and simulation?)</td><td>Implementation work was carried out at Intel(R) Core (TM) i7 CPU M60 @ 2.80 GHz in Python.</td></tr>
<tr><td>Test / Experiment How to test and characterize the design/prototype?</td><td>For the experimental work, the datasets were divided into the ratio of 80used to train classification algorithms, and the remaining 20% used for testing purposes. Accuracy, specificity, sensitivity, and area under the curve were evaluated for the seven classifiers.</td></tr>
<tr><td>Simulation/Test Data (What parameters are determined?)</td><td>Datasets was collected from - UCI repository.
<table>
<tr><th rowspan="2">No.</th><th rowspan="2">Origin</th><th rowspan="2">Destination</th><th colspan="2">Flight Type</th></tr>
<tr><th>Domestic</th><th>International</th></tr>
<tr><td>1.</td><td>Addis Ababa (ADD)</td><td>ASO (Asosa)</td><td>Yes</td><td>No</td></tr>
<tr><td>2.</td><td>Djibouti (JIB) Djibouti</td><td>Addis Ababa (ADD)</td><td>No</td><td>Yes</td></tr>
<tr><td>3.</td><td>Addis Ababa (ADD)</td><td>Entebbe (EBB) Uganda</td><td>No</td><td>Yes</td></tr>
<tr><td>4.</td><td>Hargeisa (HGA) Somalia</td><td>Addis Ababa (ADD)</td><td>No</td><td>Yes</td></tr>
<tr><td>5.</td><td>Addis Ababa (ADD)</td><td>Hawassa (AWA)</td><td>Yes</td><td>No</td></tr>
<tr><td>6.</td><td>Bahir Dar (BJR)</td><td>Addis Ababa (ADD)</td><td>Yes</td><td>No</td></tr>
<tr><td>7.</td><td>Dire Dawa (DIR)</td><td>Addis Ababa (ADD)</td><td>Yes</td><td>No</td></tr>
</table></td></tr>
<tr><td>Result / Conclusion (What was the final result?)</td><td>The proposed GRU model significantly outperforms classic machine learning models, as well as MLP and LSTM models, in terms of assessment indicators such as mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R2). The GRU model shows high accuracy in predicting air ticket prices and is considered promising for fare prediction of flight tickets. The use of the GRU model in the research addresses the limitations of traditional machine learning techniques and leverages its unique architecture to capture temporal dependencies in flight data, resulting in improved predictive performance.</td></tr>
<tr><td>Obstacles/Challenges (List the methodological obstacles if authors mentioned in the article)</td><td>The superior performance of the GRU model suggests that future work could involve further exploring and refining deep learning models for fare prediction in the civil aviation industry. The GRU model shows promise for further study in this field, indicating that future work could involve investigating and optimizing the use of GRU models for other aspects of flight fare forecasting.</td></tr>
<tr><td>Terminology (List the common basic words frequently used)</td><td>Gated Recurrent Unit (GRU),models,Mean absolute error (MAE), Root mean square error (RMSE).</td></tr>
</table>