**Title Page:**

Investigation of Aviation accident Prognosis using Logistic Regression over XG Boost algorithm to Ameliorate the accuracy

Arunkumar M1, Nelson Kennedy Babu C2

Arunkumar M1,

Research Scholar,

Department of Computer Science and Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pincode:602105.

[arunkumarm19@saveetha.com](mailto:arunkumarm19@saveetha.com)

Nelson kennedy babu C2,

Research Guide, Corresponding Author,

Department of Cloud Computing,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pincode:602105.

[nelson.sse@saveetha.com](mailto:nelson.sse@saveetha.com)

**Keywords:** Transportation, Machine Learning (ML), Novel Logistic Regression, XGBoost classifier, Aviation, Safety.

**ABSTRACT**

**Aim:** The Investigation aims to detect and identify the severity of aircraft accidents happening due to flight crew weather, maintenance, location of accident, type of engine using Machine learning (ML) approach. For this research novel Logistic regression and XGBoost classifiers were chosen and visualise the severity by comparing accuracy gain. **Materials and methods:** A dataset volume of 80% is used to train the suggested ML classifier model, while 20% is used for testing. The dataset for the present works contains 5200 unique values. The output of two classifiers is divided into two groups for SPSS analysis, with each group consis-ting of 10 output values under various functional operations, the parameters CI, alpha are valued as 0.03 (p<0.05) and G power of 0.95 is used. It shows that there is a statistical significance difference between the Novel LR algorithm and XG BOOST algorithm with p=0.002 (p<0.05). **Result:** The accuracy of the chosen Novel Logistic regression classifier, which increased recognition of indoor objects, was 95.4540%, compared to the XGBoost classifier's accuracy of 90.9770%. **Conclusion:** The proposed Novel Logistic Regression has proven to be more accurate than both the XGBoost classifier and conventional classifiers typically used for aviation accident prediction, providing evidence of the investigation's efficacy.

**Keywords:** Transportation, Machine Learning (ML), Novel Logistic Regression, XGBoost classifier, Aviation, Safety, Research.

**INTRODUCTION**

Aviation is a vital source of transportation since it offers a quick network over the globe. It promotes economic development, domestic and international trade, and tourism, as well as job possibilities and functions as a blessing in times of calamity. The airport industry has seen significant improvements since the first flight in 1903, enhancing the entire flight experience, safety rating, and expanding the network to connect all countries [(Nimmagadda, 2020a)](https://paperpile.com/c/eVbcr8/3YDx). Despite the fact that flying is often regarded as the safest means of transportation, aircraft accidents have relatively low survivability rates due to the potential for catastrophic outcomes. In the aviation business, this process of one reason often turning to another is known as the "Swiss Cheese Model" and is why there is never just one cause for an aeroplane disaster [(Srinivasan, Nagarajan, and Mahadevan 2019)](https://paperpile.com/c/eVbcr8/FlTM). So, it is crucial to take into account all factors that affect how aircraft operate. Without the use of set rules, machine learning can categorise data and generate predictions. The application of machine learning methods to forecast crash severity is the main topic of this study [(Chen, Yu, and Li 2016)](https://paperpile.com/c/eVbcr8/9af5). It has uses in the identification of plant diseases, the selection of wholesome foods, and patient health monitoring [(Chen, Yu, and Li 2016)](https://paperpile.com/c/eVbcr8/9af5).

From the year of 2001 to 2023, a grand total of 2250 articles were reviewed and the articles chosen from IEEE Xplore counts 550, 1000 from Researchgate, 500 from Google Scholar, 50 from Elsiver and 150 from Springer. [(Mathur, Khatri, and Sharma 2017)](https://paperpile.com/c/eVbcr8/b9Cc) performed analysis of aviation accident detection by experimenting with Aviation safety network dataset counts 6000 accident records. For this approach logistic regression is recommended to identify if an aeroplane was accident prone or not. For improving the prediction, the attributes namely flight crew weather, maintenance, etc. From this research, the proper maintenance of aircraft will minimise the risk factors of accidents. [(“Predicting Fatalities in Air Accidents Using CHAID XGBoost Generalized Linear Model Neural Network and Ensemble Models of Machine Learning” 2020)](https://paperpile.com/c/eVbcr8/8hjv) considered Kaggle dataset for this research and it is splitted into 3:1 ratio. The comparative analysis of aviation accident prediction using classifiers namely Neural network, Chi-squared automatic interaction detection approach, XGboost classifier, random forest and multiple linear classification technique. [(Williams 2014)](https://paperpile.com/c/eVbcr8/z9En0) diagnosed airline turbulence using random forest. [(Kaur and Kaur 2017)](https://paperpile.com/c/eVbcr8/uDSF) predict the chance of road accidents [(Kaur and Kaur 2017)](https://paperpile.com/c/eVbcr8/uDSF)and locate accident prone locations using a machine learning approach with a greater emphasis on highway and district roadways. The research done by [(Wenqi, Dongyu, and Menghua 2017)](https://paperpile.com/c/eVbcr8/1dKr) demonstrates traffic accidents identification depends on the traffic flow, climatic condition. By considering essential features TAP-CNN identified accidents. The author [(Nimmagadda, 2020b)](https://paperpile.com/c/eVbcr8/YQey) mining techniques to identify the airline accident caused due to the bird strike.

ML is a promising technique for better item identification. Using smaller datasets based on fewer parameters, previous methodologies detected and evaluated the severity of aviation accidents. This study aimed to run more accurate prediction processes inside the dataset. The greatest amount of data was utilised. This research study identifies a research gap in predicting aviation accidents using Logistic Regression against the XG Boost algorithm.

**MATERIALS AND METHODS**

This research was performed at the User Interface Laboratory in Department of Cloud Computing, Saveetha School of Engineering,  Saveetha Institute of Medical and Technical Sciences, in which the lab provides extremely superior configured systems which help to get accurate results.The sample size taken from clinical.com [(Griffin, Young, and Stanton 2015)](https://paperpile.com/c/eVbcr8/6Ic2). The dataset for the present works contains 5200 unique values. Total no.of groups considered for the research were two, group1 consists of 10 sample sizes and group 2 consists of 10 sample sizes. The calculation is done with G-power 0.82, alpha value 0.005, beta value 0.95 and confidence interval 95%.

Using a given input variable, the method of logistic regression estimates the likelihood of a discrete result.Sourced from KAGGLE's public domain [(Samaha 2023)](https://paperpile.com/c/eVbcr8/rYxb), the aviation accident prediction dataset. Proposed independent and dependent variables were employed in the experiment, and Table 1 provides a thorough overview of the dataset. Numerous traffic accidents make up the dataset, which was split into two parts: training and testing. On a laptop with an Intel i5 processor, 8GB of RAM, and the 64-bit version of Microsoft Windows 11, the project was completed using a Jupyter notebook.

**Logistic regression (LR)**

A method for calculating the likelihood of a discrete result given an input variable is logistic regression. [(Edgar and Manz 2017)](https://paperpile.com/c/eVbcr8/6wQR8). A significant portion of logistic regression models have a binary result, such as true or false, yes or no, and so on. Multinomial logistic regression can be used to model events with more than two possible outcomes. Classification problems are a good context for using logistic regression as an analytical tool when trying to discover which category a new sample most closely matches. Logistic regression is a helpful analytical method given the categorization difficulties connected to cyber security, such as threat detection.

**Pseudocode**

Dataset for Training

Output: precision

Step 1: Gather the needed volume of data.

Step 2: The following stage is pre-processing.

Step 3: Remove any noise or empty spots before proceeding with the procedure.

Step 4: Get rid of any null values.

Step 5: Identify and extract characteristics

Step 6: Use features to train the model.

Step 7: The classification process model is created and trained.

Step 8: divide the dataset into 80% for training and 20% for testing.

Step 9: The categorization has been completed with the required accuracy range.

End of Return Accuracy

**XGBoost classifier**

Gradient boosting is the foundation of the scalable ensemble technique known as XGBoost, which has been successfully used to solve ML issues. Similar to gradient boosting, XGBoost builds an additive extension of the objective function by minimising a loss function. XGBoost only employs decision trees as its fundamental classifiers, hence a separate loss function is used to control the trees' complexity [(Brownlee 2016)](https://paperpile.com/c/eVbcr8/KAHUX). With XGBoost, individual trees are constructed using a large number of cores, and data is organised to reduce lookup times. As a result, the models performed better because their training period was shorter.

**Pseudocode**

Input: Training Dataset

Output: Accuracy

Step 1: selection of dataset.

Step 2: The selected data is loaded to the network.

Step 3: The modification of dataset has to be done.

Step 4:Next stage is pre-processing.

Step 5:If any noise or empty spaces are there, it needs to be removed for further processing.

Step 6: Normalization process is done.

Step 7: Attributes are chosen and features needed for improvising the classification are extracted. Step 8: Train the model with selected features.

Step 9: Complete classification.

Return Accuracy

**Statistical Analysis**

The analysis and detection of aviation accidents on road by factors such as flight crew weather, maintenance, bird strike, traffic flow is performed. For statistical analysis of output attained from the python compiler, IBM SPSS version-26 software is utilized. The collection of previous records of aviation accidents with essential attributes is chosen as an independent variable and the goal of this study is to recognise it with higher accuracy gain. The accuray gain is considered as a dependent variable.[(George and Mallery 2019)](https://paperpile.com/c/eVbcr8/ALi6)

**RESULTS**

The detection of aircraft accidents and visualising the severity of it to ensure the safety of passengers and prevent economical loss of a nation using ML classifiers namely Novel Logistic regression and XGBoost classifier is performed. From the python compiler accuracy gain of Novel Logistic regression and XGBoost classifier recorded as 95.4540% and 90.9770%. The proposed Novel Logistic regression classifier improved the detection of objects in a room or outdoor environment effectively and its accuracy gained visualising the standard of such classifiers. Table 1 having the accuracy gain of Novel Logistic regression and XGBoost classifier obtained from python compiler at 10 different instants. The SPSS statistical analysis took data available in table 1 as input and performed a comparative mean test. The comparative mean test is categorised as group statistical analysis and independent sample test. Initially group statistics is carried out and it is denied in table 2. By taking 10 samples per group, the mean accuracy, standard deviation and standard error mean is obtained. The values obtained from group 1 are 95.4540% and group 2 having 90.9770% respectively. It shows that there is a statistical significance difference between the Novel LR algorithm and XG BOOST algorithm with p=0.002 (p<0.05).

Table 3 implies assumption and non assumption of equal variance in accuracy for selected classifiers. For this analysis, the value of p is maintained as p<0.05. The graph obtained from statistical analysis is visualised in figure 1. From table 2, the mean accuracy value is chosen and the mean accuracy comparison graph is prepared. The X-axis denotes suggested classifiers and Y-axis denotes accuracy value. The mean accuracy of proposed and conventional classifiers is 95% and 90.9770%. It shows that there is a statistical significance difference between the logistic regression and XGBoost algorithm with p = 0.02 (p<0.05).

**DISCUSSION**

The SPSS analysis performed with the outcome of group 1 and group 2 and achieved mean accuracy of 95.4540% and 90.9770% respectively. As per the experimental results, the proposed system is considered as the best approach to identify and predict aircraft accidents.

[(Christopher and Appavu 2013)](https://paperpile.com/c/eVbcr8/56r8) analyzed aviation accidents and found a solution to shorten the death rate. The decision tree classifier is applied on an FAA dataset that has 468 accident records from the year of 1970 to 2011. The proposed technique recognised aviation accidents with the best accuracy of 87.39%. Another study of [(Wu, 2019a)](https://paperpile.com/c/eVbcr8/bNsS) utilizing a random forest model to extract features and the bayesian regression performed real time aircraft accident prediction. The experimental analysis evidenced that this regression model predicted accidents with accuracy of 70.46%. [(Wu, 2019b)](https://paperpile.com/c/eVbcr8/6m8H) performed a comparative analysis of several classifiers such as KNN, SVM, random forest, gradient boosting, LR, ANN and stacking ensemble to predict the severity of aircraft accident prediction. Among that, the proposed stacking ensemble provides greater accuracy of 91.66%. [(Mehta, 2021)](https://paperpile.com/c/eVbcr8/Vjwl) performed crash analysis with XGBoost classifier and achieved higher accuracy of 86.12%. It shows that there is a statistical significance difference between the Novel LR algorithm and XG BOOST algorithm with p=0.002 (p<0.05).

Logistic regression can't be employed when there is fewer data than features because this could result in overfitting. And seems to have a linear decision surface, hence it cannot address non-linear issues. Real-world situations rarely involve linearly separable data [(Raikar, Pardeshi, and Sawale, 2023)](https://paperpile.com/c/eVbcr8/3ZvK). A customized linearly separable dataset has been drawn to proceed analysis using linear regression in future. Also utilizing the recent dataset will improvise the functionality of such classifiers.

**CONCLUSION**

The effectiveness of the overall system is crucial, and this mostly depends on how accurate the outcomes are. As this logistic regression approach has the best accuracy, it is primarily chosen for datasets from the aviation industry. The significance of feature selection and the impact of pertinent features on prediction accuracy have been the main topics of the study. The dataset has been cleared of all redundant data. The forecast helps the business and the pilot take all essential precautions to prevent an aircraft crash. Thus, the classification algorithms play a significant part in the evaluation and forecasting of data.

**DECLARATION**

**Conflict of Interest**

The authors do not have any conflict of interest associated with this manuscript.

**Author Contributions**

Author M A concerned in statistics collection, statistics analysis, manuscript, writing. Author CNKB concerned in conceptualization, statistics validation, crucial overview of manuscript.

**Acknowledgement**

The author would like to express their gratitude towards Saveetha School of Engineering, SIMATS, Saveetha University for providing the opportunities and facilities to carry out research study.

**Funding**

We are grateful to the following organization for providing financial support that enabled us to complete the study,

1. Finura Bioteks
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

**REFERENCES**

[Brownlee, Jason. 2016. *XGBoost With Python: Gradient Boosted Trees with XGBoost and Scikit-Learn*. Machine Learning Mastery.](http://paperpile.com/b/eVbcr8/KAHUX)

[Chen, Yu, Yuanlong Yu, and Ting Li. 2016. “A Vision Based Traffic Accident Detection Method Using Extreme Learning Machine.” *2016 International Conference on Advanced Robotics and Mechatronics (ICARM)*. https://doi.org/](http://paperpile.com/b/eVbcr8/9af5)[10.1109/icarm.2016.7606983](http://dx.doi.org/10.1109/icarm.2016.7606983)[.](http://paperpile.com/b/eVbcr8/9af5)

[Christopher, A. B. A., and Subramanian Appavu. 2013. “Data Mining Approaches for Aircraft Accidents Prediction: An Empirical Study on Turkey Airline.” *2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*. https://doi.org/](http://paperpile.com/b/eVbcr8/56r8)[10.1109/ice-ccn.2013.6528602](http://dx.doi.org/10.1109/ice-ccn.2013.6528602)[.](http://paperpile.com/b/eVbcr8/56r8)

[Edgar, Thomas W., and David O. Manz. 2017. *Research Methods for Cyber Security*. Syngress.](http://paperpile.com/b/eVbcr8/6wQR8)

[George, Darren, and Paul Mallery. 2019. *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference*. Routledge.](http://paperpile.com/b/eVbcr8/ALi6)

[Griffin, Thomas G. C., Mark S. Young, and Neville A. Stanton. 2015. *Human Factors Models for Aviation Accident Analysis and Prevention*. Ashgate Publishing, Ltd.](http://paperpile.com/b/eVbcr8/6Ic2)

[Kaur, Gagandeep, and Harpreet Kaur. 2017. “Prediction of the Cause of Accident and Accident Prone Location on Roads Using Data Mining Techniques.” *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. https://doi.org/](http://paperpile.com/b/eVbcr8/uDSF)[10.1109/icccnt.2017.8204001](http://dx.doi.org/10.1109/icccnt.2017.8204001)[.](http://paperpile.com/b/eVbcr8/uDSF)

[Mathur, Priyam, Sunil Kumar Khatri, and Mayank Sharma. 2017. “Prediction of Aviation Accidents Using Logistic Regression Model.” *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*. https://doi.org/](http://paperpile.com/b/eVbcr8/b9Cc)[10.1109/ictus.2017.8286102](http://dx.doi.org/10.1109/ictus.2017.8286102)[.](http://paperpile.com/b/eVbcr8/b9Cc)

[Mehta, Jay, Vaidehi Vatsaraj, Jinal Shah, and Anand Godbole. 2021. “Airplane Crash Severity Prediction Using Machine Learning.” *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. https://doi.org/](http://paperpile.com/b/eVbcr8/Vjwl)[10.1109/icccnt51525.2021.9579711](http://dx.doi.org/10.1109/icccnt51525.2021.9579711)[.](http://paperpile.com/b/eVbcr8/Vjwl)

[Nimmagadda, Sreeram, Soubraylu Sivakumar, Naveen Kumar, and D. Haritha. 2020a. “Predicting Airline Crash due to Birds Strike Using Machine Learning.” *2020 7th International Conference on Smart Structures and Systems (ICSSS)*. https://doi.org/](http://paperpile.com/b/eVbcr8/3YDx)[10.1109/icsss49621.2020.9202137](http://dx.doi.org/10.1109/icsss49621.2020.9202137)[.](http://paperpile.com/b/eVbcr8/3YDx)

[———. 2020b. “Predicting Airline Crash due to Birds Strike Using Machine Learning.” *2020 7th International Conference on Smart Structures and Systems (ICSSS)*. https://doi.org/](http://paperpile.com/b/eVbcr8/YQey)[10.1109/icsss49621.2020.9202137](http://dx.doi.org/10.1109/icsss49621.2020.9202137)[.](http://paperpile.com/b/eVbcr8/YQey)

[“Predicting Fatalities in Air Accidents Using CHAID XGBoost Generalized Linear Model Neural Network and Ensemble Models of Machine Learning.” 2020. *International Journal of Engineering and Advanced Technology*. https://doi.org/](http://paperpile.com/b/eVbcr8/8hjv)[10.35940/ijeat.c1009.0393s20](http://dx.doi.org/10.35940/ijeat.c1009.0393s20)[.](http://paperpile.com/b/eVbcr8/8hjv)

[Raikar, L. J., S. Pardeshi, and P. Sawale. “Airplane Crash Analysis and Prediction Using Machine Learning.” Accessed February 23, 2023.](http://paperpile.com/b/eVbcr8/3ZvK) <https://www.academia.edu/download/64527803/IRJET-V7I3831.pdf>[.](http://paperpile.com/b/eVbcr8/3ZvK)

[Samaha, Kheirallah. 2023. “Aviation Accident Database & Synopses, up to 2023.”](http://paperpile.com/b/eVbcr8/rYxb) <https://www.kaggle.com/khsamaha/aviation-accident-database-synopses>[.](http://paperpile.com/b/eVbcr8/rYxb)

[Srinivasan, Prabhakar, Venkataramana Nagarajan, and Sankaran Mahadevan. 2019. “Mining and Classifying Aviation Accident Reports.” *AIAA Aviation 2019 Forum*. https://doi.org/](http://paperpile.com/b/eVbcr8/FlTM)[10.2514/6.2019-2938](http://dx.doi.org/10.2514/6.2019-2938)[.](http://paperpile.com/b/eVbcr8/FlTM)

[Wenqi, Lu, Luo Dongyu, and Yan Menghua. 2017. “A Model of Traffic Accident Prediction Based on Convolutional Neural Network.” *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*. https://doi.org/](http://paperpile.com/b/eVbcr8/1dKr)[10.1109/icite.2017.8056908](http://dx.doi.org/10.1109/icite.2017.8056908)[.](http://paperpile.com/b/eVbcr8/1dKr)

[Williams, John K. 2014. “Using Random Forests to Diagnose Aviation Turbulence.” *Machine Learning* 95 (1): 51–70.](http://paperpile.com/b/eVbcr8/z9En0)

[Wu, Mingxian, Donghui Shan, Zuo Wang, Xiaoduan Sun, Jianbei Liu, and Ming Sun. 2019a. “A Bayesian Network Model for Real-Time Crash Prediction Based on Selected Variables by Random Forest.” *2019 5th International Conference on Transportation Information and Safety (ICTIS)*. https://doi.org/](http://paperpile.com/b/eVbcr8/bNsS)[10.1109/ictis.2019.8883694](http://dx.doi.org/10.1109/ictis.2019.8883694)[.](http://paperpile.com/b/eVbcr8/bNsS)

[———. 2019b. “A Bayesian Network Model for Real-Time Crash Prediction Based on Selected Variables by Random Forest.” *2019 5th International Conference on Transportation Information and Safety (ICTIS)*. https://doi.org/](http://paperpile.com/b/eVbcr8/6m8H)[10.1109/ictis.2019.8883694](http://dx.doi.org/10.1109/ictis.2019.8883694)[.](http://paperpile.com/b/eVbcr8/6m8H)

**TABLES AND FIGURE**

**Table 1.** Accuracy comparison of Logistic Regression and XGBoost classifier. The mean and standard deviation of the group and accuracy of the existing and proposed methods were 95.4540%, and 90.9770% respectively.

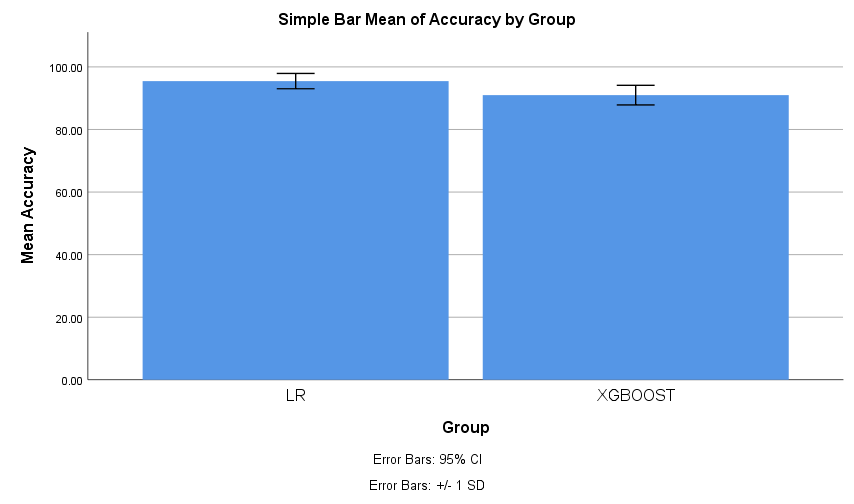
| **Logistic Regression** | **XGBoost** |
| --- | --- |
| 92.26 | 86.17 |
| 92.60 | 88.76 |
| 93.13 | 87.05 |
| 93.95 | 88.94 |
| 95.11 | 90.14 |
| 95.84 | 92.62 |
| 96.42 | 93.37 |
| 98 | 94.10 |
| 98.47 | 94.18 |
| 98.76 | 94.44 |

**Table 2.** The mean and standard deviation of the group and accuracy of the existing and proposed methods were 95.4540%, 2.44310, 90.9770%, respectively.

| **Group Statistics** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **GROUP NAME** | **N** | **Mean** | **Standard Deviation** | **Standard Error Mean** |
| **Efficiency** | **Proposed** | 10 | 95.4540 | 2.44310 | .77257 |
| **Existing** | 10 | 90.9770 | 3.13978 | .99289 |

**Table 3.** The independent sample test revealed a substantial variation in accuracy among the suggested Logistic Regression classifier and XGBoost classifier. It shows that there is a statistical significance difference between the logistic regression and XGBoost algorithm with p = 0.02 (p<0.05). It shows that there is a statistical significance difference between the Novel LR algorithm and XG BOOST algorithm with p=0.002 (p<0.05).

| **Independent sample test** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Levene’s test for equality of variance** | | | | **t-test for equality of means** | | | | | | |
| **t** | **df** | **Sig. (2-tailed)** | **Mean difference** | **Std. Error difference** | **95% confidence interval of the difference** | |
|  | | **F** | **Sig.** | **Lower** | **Upper** |
| **Efficiency** | **Equal variance assumed** | 1.932 | 0.03 | 3.559 | 18 | .002 | 4.47700 | 1.25805 | 1.83393 | 7.12007 |
| **Equal variance not assumed** |  |  | 3.559 | 16.975 | .001 | 4.47700 | 1.25805 | 1.82244 | 7.13156 |

**Fig. 1.** Bar chart representing the gain comparison between the Logistic Regression and XGBoost classifier on aviation accident prediction. X-axis represents Logistic Regression and XGBoost classifier; Y-axis represents mean accuracy ± 1SD.