**Title Page**

Aviation Accident Prediction using Logistic Regression over ADABOOST Algorithm for Better 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, ADABOOST classifier, Aviation, Safety, Research.

**ABSTRACT**

**Aim:** The proposed research uses a machine learning (ML) approach to identify and analyze aircraft accidents caused by the flight crew, weather, maintenance, accident location, and engine type during transportation. For this research novel Logistic regression and ADABOOST 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 consisting of 10 output values under various functional operations, the parameters CI, alpha are valued as 0.04 (p<0.05) and G power of 0.95 is used. It shows that there is no statistical significance difference between the Novel LR algorithm and ADABOOST algorithm with p=0.066 (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 ADABOOST classifier's accuracy of 93.4840%. **Conclusion:** The Novel Logistic Regression classifier that was suggested has a higher accuracy than the ADABOOST classifier provides accuracy. This substitutes the traditional classifiers used in aviation accident prediction and projects the effectiveness of this investigation.

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

**INTRODUCTION**

Increasingly, individuals are choosing to travel by air due to rising levels of globalisation, affordable airfares, and the safety of this form of transportation. Also, it appears to double in the upcoming years, which could increase the risk of an aviation accident [(Jeong 2021)](https://paperpile.com/c/xByP0Q/M5mU3). The current burden of transportation traffic is proving difficult for the aviation sector to handle. Due to the massive increase in aircraft transportation traffic in the restricted area, flight delays and traffic jams would get worse. Such circumstances increase the likelihood of an aviation mishap. Air Traffic Control (ATC) operators have a tremendous amount of work to do to maintain the same level of system security as before due to the increased air traffic congestion [(International Civil Aviation Organization 1984)](https://paperpile.com/c/xByP0Q/V8NkM). The research community has begun to pay attention to aircraft transportation safety because of the potentially fatal results of aviation accidents. The NTSB, ASRS, and FAA are just a few of the organisations that keep track of past aviation accident data. This previous data will be taken into account as input for the system to make safety judgements. The aim of machine learning is on algorithms capable of acquiring information from the dataset. Such forecasting methods can lessen aircraft accidents, saving numerous irreplaceable lives [(Monika. 2021)](https://paperpile.com/c/xByP0Q/CmpMN). They may also be extremely useful for planning and decision-making before takeoff. The applications are gender prediction, medical dataset theft detection and gastrointestinal disorder detection [(Kahrilas 2004)](https://paperpile.com/c/xByP0Q/irNWP).

From the year of 2007 to 2023, a grand total of 2000 articles were reviewed and the articles chosen from IEEE Xplore counts 1000, 180 from Researchgate, 170 from Elsiver and 650 from Springer. An approach for predicting aviation accidents by [(Koteeswaran. 2019)](https://paperpile.com/c/xByP0Q/M694E) combines correlation-based feature selection with the KNN algorithm. By foreseeing accident causes, this novel approach could identify hazards and enhance the aviation management system. Exploring accident causes and accident prevention are the key objectives of accident data analysis. By creating a grammar that describes states and trigger sequences, [(Rao and Marais 2020)](https://paperpile.com/c/xByP0Q/7Rd18) presented a state-based approach. The outcome demonstrates that using a rule-based approach can produce superior accident cause statistics. According to [(Hegde and Rokseth 2020)](https://paperpile.com/c/xByP0Q/q8bC2), it is possible to integrate various approaches that deal with various types of information to significantly improve the chances of preventing aviation accidents. In order to predict the age and type of aircrafts, which are two characteristics that lead to accidents, [(Altay, Ozkan, and Kayakutlu 2014)](https://paperpile.com/c/xByP0Q/VkhdV) employed genetic algorithms and ANN. [(Castilho 2015)](https://paperpile.com/c/xByP0Q/PAEb3) enhanced prediction outcomes for the detection of failure of aviation equipment by creating variables from the maintenance personnel' experience and feeding them into the Bayesian network as inputs.

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. Accuracy in predicting aviation accidents using Logistic Regression compared over ADABOOST Algorithm is a research gap that has been discovered in this work.

**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 [(Kelly and Efthymiou 2019)](https://paperpile.com/c/xByP0Q/y45w). The sample size taken from clinical.com. 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%.

The method of logistic regression is used to calculate the likelihood of a discrete outcome given an input variable. The dataset used to predict aviation accidents was taken from KAGGLE's public domain [(Samaha 2023)](https://paperpile.com/c/xByP0Q/t5JG). The experiment included proposed independent and dependent variables, and Table 1 provides a thorough overview of the dataset. Numerous traffic accidents make up the dataset, which is divided into 20% for testing and 80% for training. The project was completed using a Jupyter notebook on a laptop with an Intel i5 processor, 8GB of RAM, and the 64-bit version of Microsoft Windows 11.

**Logistic regression (LR)**

Logistic regression is an approach for estimating the probability of a discrete outcome given an input variable. [(Edgar and Manz 2017)](https://paperpile.com/c/xByP0Q/rTaPC). Most logistic regression models have a binary outcome that can be true, false, yes, or no, among other options. Regression models for events with more than two possible outcomes can be created using multinomial logistic regression. Logistic regression can be used as an analytical tool to determine which category a new sample most closely resembles when dealing with classification problems. Considering the classification difficulties related to cyber security, such as threat detection, logistic regression is a relevant analytical method.

**Algorithm**

Input : Dataset for Training

Output: precision

Step 1: Gather the required volume of data.

Step 2: The following stage is pre-processing.

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

Step 4: Get rid of any null values.

Step 5: Identify and extract features

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

**ADABOOST classifier**

An ensemble method which trains and connects trees in series combination [(Misra, Li, and He 2019)](https://paperpile.com/c/xByP0Q/bygGR). A tree is said to be stump. It implements boosting, a series connection of weak classifiers that tries to improve the classification rate of samples which was misclassified by the old weak classifier. Also the weight of trees was boosted. The term boosting converts the combination of weak classifiers into strong ones. Each tree is trained to bring attention to the weakness of the old tree.

**Algorithm**

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

End

**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/xByP0Q/6kGF).

**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 ADABOOST Classifier classifier is performed. From the python compiler accuracy gain of Novel Logistic regression and ADABOOST Classifier classifier recorded as 95.4540% and 93.4840%. 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 ADABOOST Classifier 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 93.4840% respectively.

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 95.4540% and 93.4840%. It shows that there is no statistical significance difference between the Novel LR algorithm and ADABOOST algorithm with p=0.066 (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 93.4840% respectively. As per the experimental results, the proposed system is considered as the best approach to identify and predict aircraft accidents.

In order to categorize 13,165 highway-railroad crossing occurrences, [(Soleimani. 2019)](https://paperpile.com/c/xByP0Q/oVKa5) developed the XGboost algorithm and achieved an accuracy of 99.11%. Other standard techniques, such Decision Tree or Random Forest, however, also accomplish about 98.5% accuracy. It shows that there is no statistical significance difference between the Novel LR algorithm and ADABOOST algorithm with p=0.066 (p>0.05). This follows that the incident records they were working with are probably naturally simple to distinguish. Utilizing RTA data from Ethiopia and prediction models, [(Beshah and Hill 2010)](https://paperpile.com/c/xByP0Q/cGwOl) developed an approach to examine the influence of road-related elements in accident severity. Naive Bayes, KNN, and decision tree methods are used to classify the data. Accuracy rates for decision tree and KNN were about 80%, while those for naive bayes were around 79%. The application of the decision tree approach to analyze aviation data is investigated in order to identify the best parameters for lowering the rate of aviation fatalities. The method was used to assess 468 data records from accident reports produced by the FAA between 1970 and 2011. The WEKA Decision Tree module was utilized for deployment and evaluation, and the resulting model correctly predicted accident warnings with an accuracy of 87.39% [(Christopher and alias Balamurugan 2013)](https://paperpile.com/c/xByP0Q/RLnVk). Because it is able to assess the impact of features, the Random Forest method has quite a broad range of uses in clustering as well as regression. [(Wu. 2019)](https://paperpile.com/c/xByP0Q/qMoCb) proposes a real-time crash prediction system wherein Bayesian Network Model performs the forecasting after initial Random Forest selects significant features. Sensors collected data that included fields like traffic, speed, and occupancy, which were then utilized to develop and assess the model. An ROC Curve was used to evaluate the model's performance, and it showed 70.46% accuracy.

In contrast to certain other machine learning algorithms that depend on trees, like decision trees and random forests, logistic regression is not impervious to missing data. This usually entails unnecessary work on data regarding processing missing values. Despite being less likely compared to other algorithms, logistic regression is nevertheless vulnerable to overfitting. A larger training set and regularization may be utilized to combat this tendency in future. Another illustration shows that rather than creating forecasting techniques, more attention should be paid to approach choosing and development in the future.

**CONCLUSION**

Throughout this study, a wide range of ML approaches were applied to perform a comprehensive evaluation of their impact on high-risk occurrences in aviation safety issues. The results of the simulation reveal that, in terms of recollection of the high-risk events, the stacking model performs better than the separate methods.

**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 C N K B 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. Techgrader Solutions, West Tambaram, chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

**REFERENCES**

[Altay, Ayca, Omer Ozkan, and Gulgun Kayakutlu. 2014. “Prediction of Aircraft Failure Times Using Artificial Neural Networks and Genetic Algorithms.” *Journal of Aircraft* 51 (1): 47–53.](http://paperpile.com/b/xByP0Q/VkhdV)

[Beshah, Tibebe, and Shawndra Hill. 2010. “Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia.” In *AAAI Spring Symposium: Artificial Intelligence for Development*, 24:1173–81.](http://paperpile.com/b/xByP0Q/cGwOl)

[Castilho, Isabela Xavier. 2015. “Fault Prediction in Aircraft Tires Using Bayesian Networks.” Master’s thesis, Instituto Superior Técnico, Portugal.](http://paperpile.com/b/xByP0Q/PAEb3)

[Christopher, A. B. Arockia, and S. Appavu alias Balamurugan. 2013. “Data Mining Approaches for Aircraft Accidents Prediction: An Empirical Study on Turkey Airline.” In *2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, 739–45.](http://paperpile.com/b/xByP0Q/RLnVk)

[Edgar, Thomas W., and David O. Manz. 2017. *Research Methods for Cyber Security*. Syngress.](http://paperpile.com/b/xByP0Q/rTaPC)

[George, Darren, and Paul Mallery. 2019. *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference*. Routledge.](http://paperpile.com/b/xByP0Q/6kGF)

[Hegde, Jeevith, and Børge Rokseth. 2020. “Applications of Machine Learning Methods for Engineering Risk Assessment – A Review.” *Safety Science*. https://doi.org/](http://paperpile.com/b/xByP0Q/q8bC2)[10.1016/j.ssci.2019.09.015](http://dx.doi.org/10.1016/j.ssci.2019.09.015)[.](http://paperpile.com/b/xByP0Q/q8bC2)

[International Civil Aviation Organization. 1984. *Accident Prevention Manual*.](http://paperpile.com/b/xByP0Q/V8NkM)

[Jeong, Jinwook. 2021. *An Analysis of Human Error Management on Aviation Accident Prevention*.](http://paperpile.com/b/xByP0Q/M5mU3)

[Kahrilas, P. J. 2004. “Review Article: Gastro-Oesophageal Reflux Disease as a Functional Gastrointestinal Disorder.” *Alimentary Pharmacology & Therapeutics*. https://doi.org/](http://paperpile.com/b/xByP0Q/irNWP)[10.1111/j.1365-2036.2004.02185.x](http://dx.doi.org/10.1111/j.1365-2036.2004.02185.x)[.](http://paperpile.com/b/xByP0Q/irNWP)

[Kelly, Damien, and Marina Efthymiou. 2019. “An Analysis of Human Factors in Fifty Controlled Flight into Terrain Aviation Accidents from 2007 to 2017.” *Journal of Safety Research* 69 (June): 155–65.](http://paperpile.com/b/xByP0Q/y45w)

[Koteeswaran, S., N. Malarvizhi, E. Kannan, S. Sasikala, and S. Geetha. 2019. “Data Mining Application on Aviation Accident Data for Predicting Topmost Causes for Accidents.” *Cluster Computing*. https://doi.org/](http://paperpile.com/b/xByP0Q/M694E)[10.1007/s10586-017-1394-2](http://dx.doi.org/10.1007/s10586-017-1394-2)[.](http://paperpile.com/b/xByP0Q/M694E)

[Misra, Siddharth, Hao Li, and Jiabo He. 2019. *Machine Learning for Subsurface Characterization*. Gulf Professional Publishing.](http://paperpile.com/b/xByP0Q/bygGR)

[Monika, Seema Verma, Pardeep Kumar, and Mansi Kakran. 2021. “A Stacking Model for Aviation High Risk Event Decision Making.” *Journal of Physics. Conference Series* 1950 (1): 012069.](http://paperpile.com/b/xByP0Q/CmpMN)

[Rao, Arjun H., and Karen Marais. 2020. “A State-Based Approach to Modeling General Aviation Accidents.” *Reliability Engineering & System Safety* 193 (January): 106670.](http://paperpile.com/b/xByP0Q/7Rd18)

[Samaha, Kheirallah. 2023. “Aviation Accident Database & Synopses, up to 2023.”](http://paperpile.com/b/xByP0Q/t5JG) <https://www.kaggle.com/khsamaha/aviation-accident-database-synopses>[.](http://paperpile.com/b/xByP0Q/t5JG)

[Soleimani, Samira, Saleh R. Mousa, Julius Codjoe, and Michael Leitner. 2019. “A Comprehensive Railroad-Highway Grade Crossing Consolidation Model: A Machine Learning Approach.” *Accident; Analysis and Prevention* 128 (July): 65–77.](http://paperpile.com/b/xByP0Q/oVKa5)

[Wu, Mingxian, Donghui Shan, Zuo Wang, Xiaoduan Sun, Jianbei Liu, and Ming Sun. 2019. “A Bayesian Network Model for Real-Time Crash Prediction Based on Selected Variables by Random Forest.” In *2019 5th International Conference on Transportation Information and Safety (ICTIS)*, 670–77.](http://paperpile.com/b/xByP0Q/qMoCb)

**TABLES AND FIGURE**

**Table 1.** Accuracy comparison of Logistic Regression and ADABOOST Classifier classifier. The mean and standard deviation of the group and accuracy of the existing and proposed methods were 95.4540%, and 93.4840% respectively.

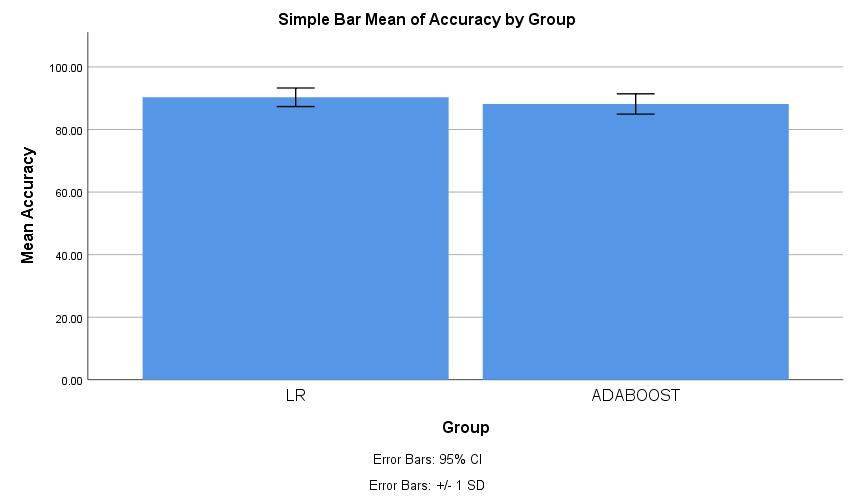
| **Logistic Regression** | **ADABOOST** |
| --- | --- |
| 92.26 | 90.26 |
| 92.60 | 91 |
| 93.13 | 92.13 |
| 93.95 | 92.95 |
| 95.11 | 93.01 |
| 95.84 | 93.84 |
| 96.42 | 94.42 |
| 98 | 95 |
| 98.47 | 95.47 |
| 98.76 | 96.76 |

**Table 2.** The mean and standard deviation of the group and accuracy of the Logistic Regression and ADABOOST Classifier algorithms were 95.4540% and 93.4840%, respectively. In comparison to the Logistic regression, ADABOOST classifier

| **Group Statistics** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **GROUP NAME** | **N** | **Mean** | **Standard Deviation** | **Standard Error Mean** |
| **Efficiency** | **Logistic Regression** | 10 | 95.4540 | 2.44310 | .77257 |
| **ADABOOST** | 10 | 93.4840 | 2.02444 | .64018 |

**Table 3.** The independent sample test revealed a substantial variation in accuracy among the suggested Logistic Regression and ADABOOST classifiers. Since p<0.05, significance level 0.04 which is less than 0.05. It shows that there is no statistical significance difference between the Novel LR algorithm and ADABOOST algorithm with p=0.066 (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** | .731 | 0.04 | 1.963 | 18 | .065 | 1.9700 | 1.00335 | -.13795 | 4.07795 |
| **Equal variance not assumed** |  |  | 1.963 | 17.399 | .066 | 1.9700 | 1.00335 | -.14318 | 4.08318 |



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