**Title page:**

Analysis of Aviation Accident Prediction using Logistic Regression over Support Vector Machine Algorithm to Enhance 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, Support Vector mechanism, Aviation, Safety, Research.

**ABSTRACT**

**Aim:** The purpose of this research is to detect and identify the severity of aircraft transportation accidents happening using Machine learning (ML) approach. For this research novel Logistic regression and Support Vector Mechanism (SVM) classifiers were chosen and visualize 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 0.03 (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 SVM algorithm with p=0.004 (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 Support Vector Mechanism classifier's accuracy of 91.3620. **Conclusion:** The proposed Novel Logistic Regression classifier demonstrates a greater level of accuracy compared to the Support Vector mechanism, surpassing traditional classifiers utilized in avaition accident prediction. This highlights the efficacy of the investigation.

**Keywords:** Transportation, Machine Learning (ML), Novel Logistic Regression, Support Vector mechanism, Aviation, Safety.

**INTRODUCTION**

Safety is essential to the aviation sector [(Shi 2022)](https://paperpile.com/c/Mv57Cz/u9MSt). The creation of routine and secure transportation operations remains difficult for the aviation sector. The number of flying accidents and incidents rises along with the complexity of aviation operations. As a result, it is still difficult to effectively retrieve and analyze aviation safety data to lower occurrences. Many things can lead to aviation incidents during transportation. The fact that numerous pertinent data columns in the Aviation Safety Reporting System database have blank values makes it more challenging to pinpoint the underlying reasons [(“Data-Driven General Aviation Aircraft Performance Modeling and Safety Research” 2020)](https://paperpile.com/c/Mv57Cz/X7oG2). A study area that models and forecasts the causes of aviation disasters has arisen in recent decades. Recent studies have started utilizing big data, including deep learning and data stream approaches, to forecast dangers in the aviation sector [(Луппо and Алєксєєв 2007)](https://paperpile.com/c/Mv57Cz/c0Duh). Some of the applications of it are listed below: tongue tumor detection, fertilizer recommendation, credit card fraud detection [(Kimura 2009)](https://paperpile.com/c/Mv57Cz/aPyLw).

From the year of 2007 to 2023, a grand total of 1800 articles were reviewed and the articles chosen from IEEE Xplore counts 700, 280 from Researchgate, 620 from Google Scholar, 70 from Elsiver and 130 from Springer. [(Burnett, Alan Burnett, and Si 2017)](https://paperpile.com/c/Mv57Cz/CVbiU) predict the fatalities based on classifiers namely decision tree, KNN, SVM, ANN based on the FAA data. In this research redundant parameters were eliminated and categorized numerical value. The ANN classifier outperformed several other classifiers in predicting the number of injuries and transportation casualties. [(Lee 2019)](https://paperpile.com/c/Mv57Cz/vHEqf) proposed a real time anomalous flight behavior. Decimation is used to pre-process the raw data in order to remove high frequency components. Moreover, the necessary features were retrieved, and the SVR model was trained to recognise the state of the engine. [(Valdés.2018)](https://paperpile.com/c/Mv57Cz/Prdvs) explored safety incident data, detect anomalies, access the risk range, develop an objective system for contrasting air carriers, and finally identify and anticipate incidents. For this analysis bayesian inference and hierarchical structure is proposed. [(Zhou. 2020)](https://paperpile.com/c/Mv57Cz/5RK2b) proposed a hybrid model to detect and predict civil aircraft hazard based on the ML classifiers such as PSO, SVM and LSTM classifier. In addition to conducting an innovative examination, [(A. B. Arockia Christopher and alias Balamurugan 2013)](https://paperpile.com/c/Mv57Cz/P3cDh) characterized and discussed several bunching and characterization techniques for estimating the level of warning for airplane accidents. This study primarily looked at KNN, naive bayes, decision trees, neural networks, and other techniques for forecasting the warning level in aircraft accidents.

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 utilized. The research done here is to predict aircraft accidents accurately is the area of research that has to be filled, according to this paper. The research gap mentioned in this work is a lack of accuracy in forecasting aircraft accidents using Logistic Regression compared to the Support Vector Machine 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 [(Wiegmann and Shappell 2017)](https://paperpile.com/c/Mv57Cz/KK9V).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 aviation accident prediction dataset was sourced from KAGGLE's public domain [(Samaha 2023)](https://paperpile.com/c/Mv57Cz/dNUl). Proposed independent and dependent variables were used in the experiment, and a comprehensive summary of the dataset is provided in Table 1. The dataset comprises numerous traffic accidents, with 80% of the data used for training and the remaining 20% for testing. The project was carried out using a Jupyter notebook on a laptop equipped with a Intel i5 processor, 8GB RAM, and a 64-bit Microsoft Windows 11 operating system.

**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/Mv57Cz/Qwypt). A binary result, such as true or false, yes or no, and so forth, is present in the majority of logistic regression models. It is feasible to model events with more than two possible outcomes using multinomial logistic regression. Classification problems are a suitable place to use logistic regression as an analysis tool when trying to determine which category a new sample most closely matches. Given the categorization challenges associated with cyber security, such as threat detection, logistic regression is a useful analytical technique.

**Algorithm**

Input: Training Dataset

Output: Accuracy

Step 1: get the necessary amount of data.

Step 2: Pre-processing comes next.

Step 3: In order to continue processing, any noise or empty spots must be eliminated.

Step 4: Remove null values in step four.

Step 5: Extract features in step five.

Step 6: Train the model with features in step six.

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

Step 8: 80% of the dataset will be used for training and the remaining 20% for testing.

Step 9: The classification is completed with the necessary level of accuracy.

Return Accuracy

End

**SVM**

A supervised learning technique used to do classification and regression A set of objects were trained and classified into multiple classes to identify hyperplanes in the data space which creates a greater minimum distance known as margin in between objects and classes. Then the hyperplane in aviation transportation is called the maximum margin hyperplane. It uses objects on the corners of the margin to split objects without utilizing variation among class means. The hyperplane used for separation is supported by vectors nearer the margin; so it is called a Support vector mechanism (SVM).

**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/Mv57Cz/7cRY)**.**

**RESULTS**

The detection of aircraft accidents and visualizing 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 Support vector mechanism classifier is performed. From the python compiler accuracy gain of Novel Logistic regression and Support vector mechanism classifier recorded as 95.4540 and 91.3620%. The proposed Novel Logistic regression classifier improved the detection of objects in a room or outdoor environment effectively and its accuracy gained visualizing the standard of such classifiers. It shows that there is a statistical significance difference between the Novel LR algorithm and SVM algorithm with p=0.004 (p<0.05).

Table 1 having the accuracy gain of Novel Logistic regression and Support vector mechanism 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 categorized as group statistical analysis and independent sample test. First group statistics is carried out and it is denied in table 2. The mean accuracy, standard deviation, and standard error mean are calculated using 10 samples per group. The values obtained from group 1 95.4540 and group 2 having 91.3620% 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 visualized 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 91.3620

**DISCUSSION**

The SPSS analysis performed with the outcome of group 1 and group 2 and achieved mean accuracy of 95.4540% and 91.3620% respectively. As per the experimental results, the proposed system is considered as the best approach to identify and predict aircraft accidents. It shows that there is a statistical significance difference between the Novel LR algorithm and SVM algorithm with p=0.004 (p<0.05).

[(Alkhamisi and Mehmood 2020)](https://paperpile.com/c/Mv57Cz/COYTd) combined ensemble machine learning and deep learning approach to identify the risk level to ensure aviation system security. The aviation reporting system dataset is selected for this analysis which possesses details of aviation data for a duration of 12 years. The suggested SVM and RNN fused and developed a risk prediction technique which attained overall better accuracy. A prediction model for aviation accidents and incidents was developed by [(Burnett, Alan Burnett, and Si 2017)](https://paperpile.com/c/Mv57Cz/CVbiU), and it forecasts that the majority of accidents are caused by people. This researcher used SVM, and the application obtained an improved accuracy in this investigation. According to this study by [(Bazargan and Guzhva 2011)](https://paperpile.com/c/Mv57Cz/iDU2G), only a few attributes were taken into account for implementation, and the model that was utilized was Logistic Regression, that was implemented utilizing statistical analysis tools. A 88% accuracy rate was reached by the author. [(A. B. A. Christopher and Appavu 2013)](https://paperpile.com/c/Mv57Cz/HaXPz) uses a Decision Tree algorithm to explore incidents involving Turkey Airlines using data retrieved from the FAA database over a 41-year period. with a successful accuracy rate.

The logistic regression is sensitive to outliers. The effectiveness of the model can be affected by the presence of multicollinearity among independent variables. Neural networks, which use more advanced algorithms like random forests, might actually do better than logistic regression. The fusion of several other classifiers like LSTM, neural network, SVM with logistic regression model will improvise the prediction in future and it overcomes the drawbacks as earlier discussed. Additional idea to be taken into account in order to comprehend how this approach may perform in a scaled situation is the inclusion of a larger labeled and unlabelled data collection.

**CONCLUSION**

The findings imply that cutting-edge analysis techniques may be utilized to make predictions about the seriousness of upcoming accidents and make use of previous accident data. also show how machine learning techniques can produce insightful data which goes beyond what is possible with conventional statistical analysis techniques.

**DECLARATION**

**Conflict of Interest**

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

**Author Contributions**

Author A K concerned in statistics collection, statistics analysis, manuscript, writing. Author CNKB is concerned with conceptualization, statistics validation, and 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 thank the following organization for providing financial support that enabled us to complete the study

1. Techgrader Solutions.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

**REFERENCES**

[Alkhamisi, Abrar Omar, and Rashid Mehmood. 2020. “An Ensemble Machine and Deep Learning Model for Risk Prediction in Aviation Systems.” *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*. https://doi.org/](http://paperpile.com/b/Mv57Cz/COYTd)[10.1109/cdma47397.2020.00015](http://dx.doi.org/10.1109/cdma47397.2020.00015)[.](http://paperpile.com/b/Mv57Cz/COYTd)

[Bazargan, Massoud, and Vitaly S. Guzhva. 2011. “Impact of Gender, Age and Experience of Pilots on General Aviation Accidents.” *Accident; Analysis and Prevention* 43 (3): 962–70.](http://paperpile.com/b/Mv57Cz/iDU2G)

[Burnett, R. Alan, R. Alan Burnett, and Dong Si. 2017. “Prediction of Injuries and Fatalities in Aviation Accidents through Machine Learning.” *Proceedings of the International Conference on Compute and Data Analysis*. https://doi.org/](http://paperpile.com/b/Mv57Cz/CVbiU)[10.1145/3093241.3093288](http://dx.doi.org/10.1145/3093241.3093288)[.](http://paperpile.com/b/Mv57Cz/CVbiU)

[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/Mv57Cz/HaXPz)[10.1109/ice-ccn.2013.6528602](http://dx.doi.org/10.1109/ice-ccn.2013.6528602)[.](http://paperpile.com/b/Mv57Cz/HaXPz)

[Christopher, A. B. Arockia, and S. Appavu alias Balamurugan. 2013. “Performance of Different Clustering Methods and Classification Algorithms for Prediction of Warning Level in Aircraft Accidents: An Empirical Study.” In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 1–7.](http://paperpile.com/b/Mv57Cz/P3cDh)

[“Data-Driven General Aviation Aircraft Performance Modeling and Safety Research.” 2020. https://doi.org/](http://paperpile.com/b/Mv57Cz/X7oG2)[10.2514/6.2020-3097.vid](http://dx.doi.org/10.2514/6.2020-3097.vid)[.](http://paperpile.com/b/Mv57Cz/X7oG2)

[Edgar, Thomas W., and David O. Manz. 2017. *Research Methods for Cyber Security*. Syngress.](http://paperpile.com/b/Mv57Cz/Qwypt)

[George, Darren, and Paul Mallery. 2019. *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference*. Routledge.](http://paperpile.com/b/Mv57Cz/7cRY)

[Kimura, Masaomi. 2009. “Application of Data Mining Techniques to the Data Analyses to Ensure Safety of Medicine Usage.” *Data Mining and Knowledge Discovery in Real Life Applications*. https://doi.org/](http://paperpile.com/b/Mv57Cz/aPyLw)[10.5772/6455](http://dx.doi.org/10.5772/6455)[.](http://paperpile.com/b/Mv57Cz/aPyLw)

[Lee, Hyunseong, Guoyi Li, Ashwin Rai, and Aditi Chattopadhyay. 2019. “Anomaly Detection of Aircraft System Using Kernel-Based Learning Algorithm.” *AIAA Scitech 2019 Forum*. https://doi.org/](http://paperpile.com/b/Mv57Cz/vHEqf)[10.2514/6.2019-1224](http://dx.doi.org/10.2514/6.2019-1224)[.](http://paperpile.com/b/Mv57Cz/vHEqf)

[Samaha, Kheirallah. 2023. “Aviation Accident Database & Synopses, up to 2023.”](http://paperpile.com/b/Mv57Cz/dNUl) <https://www.kaggle.com/khsamaha/aviation-accident-database-synopses>[.](http://paperpile.com/b/Mv57Cz/dNUl)

[Shi, Donghui, Shuai Cao, Jozef Zurada, and Jian Guan. 2022. “An Innovative Approach to Modeling Aviation Safety Incidents.” *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/](http://paperpile.com/b/Mv57Cz/u9MSt)[10.24251/hicss.2022.150](http://dx.doi.org/10.24251/hicss.2022.150)[.](http://paperpile.com/b/Mv57Cz/u9MSt)

[Valdés, Rosa María Arnaldo, V. Fernando Gómez Comendador, Luis Perez Sanz, and Alvaro Rodriguez Sanz. 2018. “Prediction of Aircraft Safety Incidents Using Bayesian Inference and Hierarchical Structures.” *Safety Science*. https://doi.org/](http://paperpile.com/b/Mv57Cz/Prdvs)[10.1016/j.ssci.2018.01.008](http://dx.doi.org/10.1016/j.ssci.2018.01.008)[.](http://paperpile.com/b/Mv57Cz/Prdvs)

[Wiegmann, Douglas A., and Scott A. Shappell. 2017. *A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System*. Routledge.](http://paperpile.com/b/Mv57Cz/KK9V)

[Zhou, Di, Xiao Zhuang, Hongfu Zuo, Han Wang, and Hongsheng Yan. 2020. “Deep Learning-Based Approach for Civil Aircraft Hazard Identification and Prediction.” *IEEE Access* 8: 103665–83.](http://paperpile.com/b/Mv57Cz/5RK2b)

[Луппо, O. Є., and О. М. Алєксєєв. 2007. “New Approaches to the Classification of Causes of Aviation Accidents and Incidents as an Effective Means of Improving Safety.” *Proceedings of National Aviation University*. https://doi.org/](http://paperpile.com/b/Mv57Cz/c0Duh)[10.18372/2306-1472.33.1490](http://dx.doi.org/10.18372/2306-1472.33.1490)[.](http://paperpile.com/b/Mv57Cz/c0Duh)

**TABLES AND FIGURE**

**Table 1.** Accuracy comparison of Logistic Regression and Support vector mechanism classifier. The mean and standard deviation of the group and accuracy of the existing and proposed methods were 95.4540, and 91.3620 respectively.

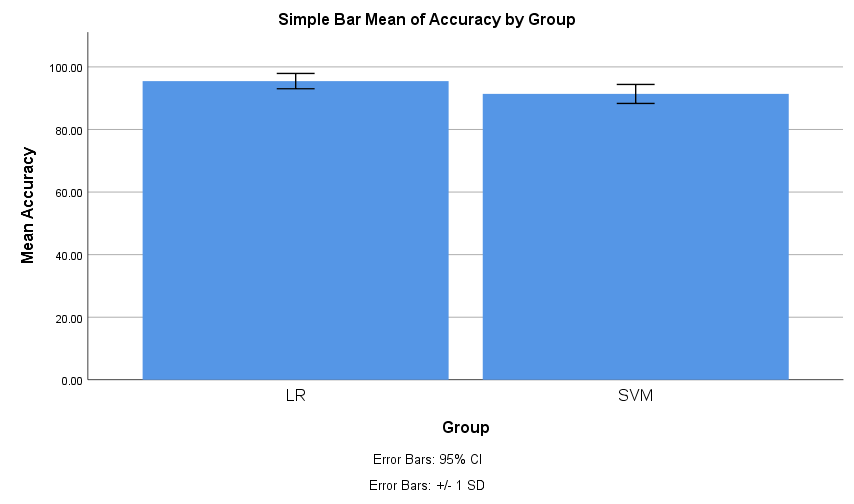
| **Accuracy (%)** | |
| --- | --- |
| **Logistic Regression** | **Support vector mechanism** |
| 92.26 | 86.17 |
| 92.60 | 87.76 |
| 93.13 | 89.05 |
| 93.95 | 89.94 |
| 95.11 | 91.14 |
| 95.84 | 92.62 |
| 96.42 | 93.37 |
| 98.00 | 94.18 |
| 98.47 | 94.44 |
| 98.76 | 94.95 |

**Table 2.** The mean and standard deviation of the group and accuracy of the Logistic Regression 95.4540 and Support vector mechanism algorithms were 91.3620 respectively. In comparison to the Support vector mechanism.

| **Group Statistics** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **GROUP NAME** | **N** | **Mean** | **Standard Deviation** | **Standard Error Mean** |
| **Accuracy** | **Logistic Regression** | 10 | 95.4540 | 2.44310 | .77257 |
| **Support vector mechanism** | 10 | 91.3620 | 3.04141 | .96178 |

**Table 3.** The independent sample test revealed a substantial variation in accuracy among the suggested Logistic Regression and Support vector mechanism classifiers. Since p<0.05, there is a substantial variation among two methods. It shows that there is a statistical significance difference between the Novel LR algorithm and SVM algorithm with p=0.004 (p<0.05).

| **Independent Sample Test** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Levene’s Test for Equality of Variances** | | | | **T-test for Equality of Means** | | | | | | |
|  | | **F** | **Sig.** | **T** | **Df** | **Sig. (2-tailed)** | **Mean Difference** | **Std. Error Differences** | **95% Confidence Interval of the Difference** | |
| **Lower** | **Upper** |
| **Accuracy** | **Equal Variances assumed** | .764 | 0.03 | 3.317 | 18 | .004 | 4.09200 | 1.23365 | 1.50020 | 6.68380 |
| **Equal Variances not assumed** |  |  | 3.317 | 17.200 | .004 | 4.09200 | 1.23365 | 1.49154 | 6.69246 |

****

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