**Title page**

Predictive Analysis of Aviation Accident using Logistic Regression over K-Nearest Neighbors Algorithm to Improve 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.

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

**Keywords:** Transportation, Machine Learning (ML), Novel Logistic Regression, K-Nearest Neighboring, Aviation, Safety, Research.

**ABSTRACT**

**Aim:** The aim of the study to attain accuracy in aviation accident prediction using Novel Logistic regression compared with K-Nearest Neighbors algorithm . **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, alpha are valued as 0.02 (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 KNN algorithm with p=0.025 (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 K-nearest neighboring classifier's accuracy of 92.7840. **Conclusion:** The investigation's efficacy is evidenced by the higher level of accuracy demonstrated by the proposed Novel Logistic Regression when compared to both the K-nearest neighboring classifier and traditional classifiers typically employed in aviation accident prediction.

**Keywords:** Transportation, Machine Learning (ML), Novel Logistic Regression, K-nearest neighboring, Aviation, Safety, Research.

**INTRODUCTION**

Air transportation plays a significant role in providing quick and secure access to locations around the globe [(Kuşkapan, Sahraei̇, and Çodur)](https://paperpile.com/c/AaBtKT/RNdTb). Particularly for quick transportation access during long distance safety travel, it is highly favoured. When air transportation initially began, it was extremely risky for passengers' safety, but as science advances, it became more dependable. But, as more flights are taken, the number of aviation mishaps has significantly increased. The number of deaths and injuries is also high as the majority of these accidents happen at speeds faster than the speed of sound. If the planes crash in this scenario from a specific height, it is doubtful that anyone will escape. It is vital to categorise them and carry out in-depth study pertaining to the given classifications in order to decrease these life losses. About 75% of aeroplane accidents and incidents are caused by human error. In addition to the human aspect, weather circumstances, bird strikes, cabin pressure issues, technical issues, and fuel issues can all contribute to aviation accidents [(Oster, Strong, and Zorn)](https://paperpile.com/c/AaBtKT/7iTiL). Production and maintenance issues with the aircraft may result in technical flaws. It is possible to analyse factors impacting plane crashes using a variety of methodologies in an effort to prevent aviation mishaps. The development of the systems used for forecasting and evaluating historical data is a result of the expansion in technology. With the aid of machine learning, a system can be made to learn automatically without explicit user input [(Raikar, Pardeshi, and Sawale, 2023)](https://paperpile.com/c/AaBtKT/Q69jA). The applications of it are bitcoin price prediction, brain tumor segmentation and heart disease prediction [(Nisbet, Miner, and Yale 2017)](https://paperpile.com/c/AaBtKT/6Nuyh).

From the year of 2011 to 2023, a grand total of 1800 articles were reviewed and the articles chosen from IEEE Xplore counts 940, 620 from Researchgate, 100 from Elsiver and 140 from Springer. The researchers [(Diamoutene, 2018)](https://paperpile.com/c/AaBtKT/19Q8j) offer an analysis on accidents using statistical tools. The flight's goal, the characteristics of the aircraft, and the location were the factors taken into account. The Gaussian Navie Bayes, Decision tree, and KNN classifier are just a handful of the many diverse models that were used. A thorough investigation was conducted on aircraft loss of control. According to [(Ancel, 2015)](https://paperpile.com/c/AaBtKT/NXwHo), a fairly straightforward model was created that allowed for the use of NASA AvSP to check numerous implications on such accidents. The analysis was based on information gathered from the NTSB over a 22-year period. By educating the cabin crew and ground workers, the airline hierarchy would be able to resolve the problem. According to one of these studies, [(Boyd 2016)](https://paperpile.com/c/AaBtKT/xjO7R), weight excess would lead to accidents. A plane's ability to climb a gradient is constrained by gravity. This is due to the distinct methodology used in the airframe's design. Also, the calculation of the Poisson Distribution, T-test, and Percentage test was done using statistical methods. Machine learning ensemble models were utilised to anticipate the incidences. Ensemble of neural networks and hybrid SVM were created by [(Zhang and Mahadevan 2019)](https://paperpile.com/c/AaBtKT/6hWPo) to address the reasons and risk of technical failure in aircrafts. [(Handel and Yackel 2011)](https://paperpile.com/c/AaBtKT/9RgtH) employ a Multiple Logistic Regression approach for a comparable analysis that includes occurrences and accidents related to commercial, medical, and ground aircraft. predicting fixed wing 2-4 capacity aircraft based on data gathered between 1984 and 2009.

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. The research gap identified here in this work is a shortfall in accuracy in predicting aviation accidents Logistic Regression compared over K-Nearest Neighbors 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 [(Griffioen 2011)](https://paperpile.com/c/AaBtKT/a5Tb). 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%.

A method for calculating the likelihood of a discrete result given an input variable is logistic regression.The public domain of KAGGLE served as the source for the aviation accident prediction dataset [(Samaha 2023)](https://paperpile.com/c/AaBtKT/kE7b). In the experiment, proposed independent and dependent variables were employed. Table 1 gives a detailed overview of the dataset. Numerous traffic incidents make up the dataset, with 80% of the data being utilized for training and the remaining 20% for testing. The project was completed using a laptop running a 64-bit version of Microsoft Windows 11 with an Intel i5 CPU, 8GB of RAM, and a Jupyter notebook.

**Logistic regression (LR)**

Logistic regression is a technique for estimating the probability of a discrete result given an input variable [(Edgar and Manz 2017)](https://paperpile.com/c/AaBtKT/udyko). A binary outcome, such as true or false, yes or no, and so forth, is present in the majority of logistic regression models. It is possible to model events with more than two possible outcomes using multinomial logistic regression. When attempting to determine which category a new sample most closely resembles, classification problems are an appropriate setting in which to use logistic regression as an analytical tool. Logistic regression is a helpful analytical method given the categorization difficulties connected to cyber security, such as threat detection.

**Algorithm**

Input: Training Dataset

Output: Accuracy

Step 1: Collecting required volume of dataset.

Step 2: Next stage is pre-processing.

Step 3: If any noise or empty spaces are there, it needs to be removed for further processing.

Step 4: Remove null values.

Step 5: extract features  
Step 6: train the model with features

Step 7: The model for the classification process is developed and trained.  
 Step 8: allocating 80% of the dataset for training and remaining 20% for testing.

Step 9: The classification is done with required accuracy range.

Return Accuracy

End

**KNN**

The supervised learning technique underlies the simplest ML classifier [(Rajaguru and Prabhakar 2017)](https://paperpile.com/c/AaBtKT/Co1uV). This nonparametric method's principal use is in classification. In order to identify the class of the new data, the classes of the K nearest data are compared with the known data, which is retained under the names of the chosen characteristics. The use of large datasets for training is common. It is widely used for making graphs [(Chanal, 2022)](https://paperpile.com/c/AaBtKT/zfzyt).

**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/AaBtKT/l3wb).

**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 K-Nearest Neighboring classifier is performed. From the python compiler accuracy gain of Novel Logistic regression and K-Nearest Neighboring classifier recorded as 95.454% and 92.784%. 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.

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. Table 1 presents the accuracy increase of Novel Logistic Regression and K-Nearest Neighbors classifier as obtained from the Python compiler over 10 different time points. 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 95.454% and group 2 having 92.784%, 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 95.454% and 92.784%. It shows that there is a statistical significance difference between the Novel LR algorithm and KNN algorithm with p=0.025 (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 92.7840% respectively. As per the experimental results, the proposed system is considered as the best approach to identify and predict aircraft accidents.

[(Sameen and Pradhan 2017)](https://paperpile.com/c/AaBtKT/8uJJW) 1130 accident reports collected from 2009 to 2015 on the North-South Expressway (NSE) in Malaysia to train a Recurrent Neural Network (RNN) to predict accident severity. The authors contrast their suggested RNN model with Bayesian Logistic Regression and Multilayer Perceptron. The RNN model's validation accuracy was 71.77%, compared to the MLP and BLR models' 65.48% and 58.30%, respectively. [(Lv, 2010)](https://paperpile.com/c/AaBtKT/hOs08) proposed a Backpropagation Neural Network-based intersection accident prediction system that draws a link among accident types and their causes. The accuracy of the network, which was trained utilizing 197 intersection data, was up to 89%. Deeper anomaly identification was conducted by [(Pusadan, Buliali, and Ginardi 2019)](https://paperpile.com/c/AaBtKT/UJQxG) depending on segment construction and testing from cluster distance, yielding 96% K-NN and K-Means accuracy and 93% SVM. [(Sun, Ellerbroek, and Hoekstra 2017)](https://paperpile.com/c/AaBtKT/bPhYa) presents the identification of big and dispersed aircraft trajectories in the ADS-B flight data with the DBSCAN approach to test for odd outliers, yielding a 95.454% detection accuracy. It shows that there is a statistical significance difference between the Novel LR algorithm and KNN algorithm with p=0.025 (p<0.05).

A large dataset and enough training examples are needed for all the categories that Logistic Regression needs to detect. Every training example must be separate from every other sample in the dataset. The model will attempt to prioritize those particular training instances if they are somehow connected. Thus, matching data or repeated measurements shouldn't be used to generate the training data. For instance, certain scientific study methods make use of several observations of the same subjects. In such circumstances, this method is inapplicable. By utilizing present dataset with large amounts of features will attempt to improvise the prediction accuracy of suggested models in future. To further increase the model's accuracy in the future, flight schedule information ought to be gathered.

**CONCLUSION**

This investigation used K-nearest neighbours and logistic regression algorithm to classify aviation accidents and mortality rates. The research identifies the method that produced the best results for investigations using comparable data sets. With the use of this method, analyses can produce more precise results. Researchers and readers can use the findings from this investigation to determine which method to use for a similar dataset in the future.

**DECLARATION**

**Conflict of Interest**

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

**Author Contributions**

Author AK 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 thank 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**

[Ancel, Ersin, Ann T. Shih, Sharon M. Jones, Mary S. Reveley, James T. Luxhøj, and Joni K. Evans. 2015. “Predictive Safety Analytics: Inferring Aviation Accident Shaping Factors and Causation.” *Journal of Risk Research*. https://doi.org/](http://paperpile.com/b/AaBtKT/NXwHo)[10.1080/13669877.2014.896402](http://dx.doi.org/10.1080/13669877.2014.896402)[.](http://paperpile.com/b/AaBtKT/NXwHo)

[Boyd, Douglas D. 2016. “General Aviation Accidents Related to Exceedance of Airplane Weight/center of Gravity Limits.” *Accident; Analysis and Prevention* 91 (June): 19–23.](http://paperpile.com/b/AaBtKT/xjO7R)

[Chanal, Damien, Nadia Yousfi Steiner, Raffaele Petrone, Didier Chamagne, and Marie-Cécile Péra. 2022. “Online Diagnosis of PEM Fuel Cell by Fuzzy C-Means Clustering.” *Encyclopedia of Energy Storage*. https://doi.org/](http://paperpile.com/b/AaBtKT/zfzyt)[10.1016/b978-0-12-819723-3.00099-8](http://dx.doi.org/10.1016/b978-0-12-819723-3.00099-8)[.](http://paperpile.com/b/AaBtKT/zfzyt)

[Diamoutene, Abdoulaye, Bernard Kamsu-Foguem, Farid Noureddine, and Diakarya Barro. 2018. “Prediction of U.S. General Aviation Fatalities from Extreme Value Approach.” *Transportation Research Part A: Policy and Practice*. https://doi.org/](http://paperpile.com/b/AaBtKT/19Q8j)[10.1016/j.tra.2018.01.022](http://dx.doi.org/10.1016/j.tra.2018.01.022)[.](http://paperpile.com/b/AaBtKT/19Q8j)

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

[George, Darren, and Paul Mallery. 2019. *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference*. Routledge.](http://paperpile.com/b/AaBtKT/l3wb)

[Griffioen, Editor Hans. 2011. *Air Crash Investigations: The Deadliest Single Aircraft Accident in Aviation History the Crash of Japan Airlines Flight 123*. Lulu.com.](http://paperpile.com/b/AaBtKT/a5Tb)

[Handel, Daniel A., and Thomas R. Yackel. 2011. “Fixed-Wing Medical Transport Crashes: Characteristics Associated with Fatal Outcomes.” *Air Medical Journal* 30 (3): 149–52.](http://paperpile.com/b/AaBtKT/9RgtH)

[Kuşkapan, Emre, Mohammad Ali Sahraei̇, and Muhammed Yasin Çodur. “Classification of Aviation Accidents Using Data Mining Algorithms.” *Balkan Journal of Electrical and Computer Engineering* 10 (1): 10–15.](http://paperpile.com/b/AaBtKT/RNdTb)

[Lv, Yuejing, Zhang Haixia, Zhou Xing-lin, Liu Ming, and Li Jie. 2010. “Research on Accident Prediction of Intersection and Identification Method of Prominent Accident Form Based on Back Propagation Neural Network.” *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*. https://doi.org/](http://paperpile.com/b/AaBtKT/hOs08)[10.1109/iccasm.2010.5620615](http://dx.doi.org/10.1109/iccasm.2010.5620615)[.](http://paperpile.com/b/AaBtKT/hOs08)

[Nisbet, Robert, Gary Miner, and Ken Yale. 2017. *Handbook of Statistical Analysis and Data Mining Applications*. Elsevier.](http://paperpile.com/b/AaBtKT/6Nuyh)

[Oster, C. V., J. S. Strong, and C. K. Zorn. “Investigation of Accidents Related to Air Traffic Control.” *51st Annual Transportation Research Forum*.](http://paperpile.com/b/AaBtKT/7iTiL)

[Pusadan, Mohammad Yazdi, Joko Lianto Buliali, and Raden Venantius Hari Ginardi. 2019. “Cluster Phenomenon to Determine Anomaly Detection of Flight Route.” *Procedia Computer Science*. https://doi.org/](http://paperpile.com/b/AaBtKT/UJQxG)[10.1016/j.procs.2019.11.151](http://dx.doi.org/10.1016/j.procs.2019.11.151)[.](http://paperpile.com/b/AaBtKT/UJQxG)

[Raikar, L. J., S. Pardeshi, and P. Sawale. “Airplane Crash Analysis and Prediction Using Machine Learning.” Accessed February 23, 2023.](http://paperpile.com/b/AaBtKT/Q69jA) <https://www.academia.edu/download/64527803/IRJET-V7I3831.pdf>[.](http://paperpile.com/b/AaBtKT/Q69jA)

[Rajaguru, Harikumar, and Sunil Kumar Prabhakar. 2017. *KNN Classifier and K-Means Clustering for Robust Classification of Epilepsy from EEG Signals. A Detailed Analysis*. Anchor Academic Publishing.](http://paperpile.com/b/AaBtKT/Co1uV)

[Samaha, Kheirallah. 2023. “Aviation Accident Database & Synopses, up to 2023.”](http://paperpile.com/b/AaBtKT/kE7b) <https://www.kaggle.com/khsamaha/aviation-accident-database-synopses>[.](http://paperpile.com/b/AaBtKT/kE7b)

[Sameen, Maher Ibrahim, and Biswajeet Pradhan. 2017. “Severity Prediction of Traffic Accidents with Recurrent Neural Networks.” *NATO Advanced Science Institutes Series E: Applied Sciences* 7 (6): 476.](http://paperpile.com/b/AaBtKT/8uJJW)

[Sun, Junzi, Joost Ellerbroek, and Jacco Hoekstra. 2017. “Flight Extraction and Phase Identification for Large Automatic Dependent Surveillance–Broadcast Datasets.” *Journal of Aerospace Information Systems* 14 (10): 566–72.](http://paperpile.com/b/AaBtKT/bPhYa)

[Zhang, Xiaoge, and Sankaran Mahadevan. 2019. “Ensemble Machine Learning Models for Aviation Incident Risk Prediction.” *Decision Support Systems*. https://doi.org/](http://paperpile.com/b/AaBtKT/6hWPo)[10.1016/j.dss.2018.10.009](http://dx.doi.org/10.1016/j.dss.2018.10.009)[.](http://paperpile.com/b/AaBtKT/6hWPo)

**TABLES AND FIGURE**

**Table 1.** Accuracy comparison of Logistic Regression and K-Nearest Neighboring classifier. The mean and standard deviation of the group and accuracy of the existing and proposed methods were 95.454%, and 92.784% respectively.

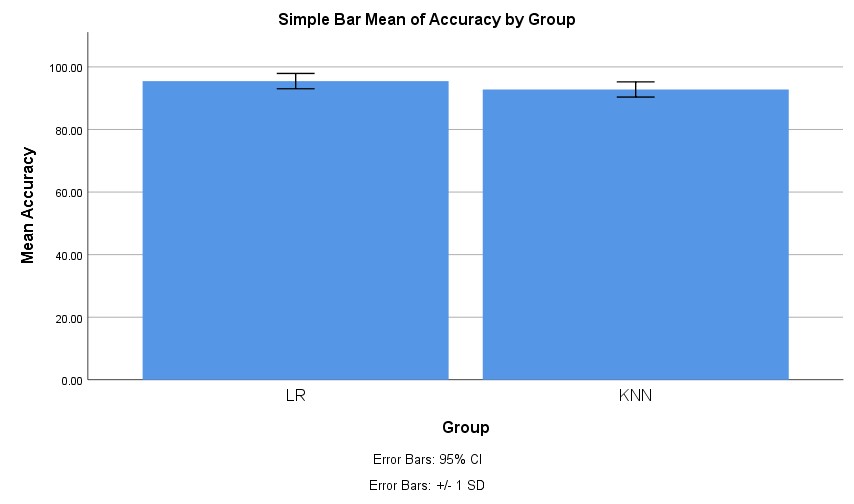
| **Accuracy (%)** | |
| --- | --- |
| **Logistic Regression** | **K-Nearest Neighboring** |
| 92.26 | 89.26 |
| 92.60 | 90.00 |
| 93.13 | 90.13 |
| 93.95 | 91.95 |
| 95.11 | 92.01 |
| 95.84 | 93.84 |
| 96.42 | 94.42 |
| 98 | 95.00 |
| 98.47 | 95.47 |
| 98.76 | 95.76 |

**Table 2.** The mean and standard deviation of the group and accuracy of the Logistic Regression and K-Nearest Neighboring algorithms were 95.454%and 92% and respectively. In comparison to the K-Nearest Neighboring, Logistic Regression.

| **Group Statistics** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **GROUP NAME** | **N** | **Mean** | **Standard Deviation** | **Standard Error Mean** |
| **Accuracy** | **Logistic Regression** | 10 | 95.4540 | 2.44310 | .77257 |
| **K-Nearest Neighboring** | 10 | 92.7840 | 2.43313 | .76942 |

**Table 3.** The independent sample test revealed a substantial variation in accuracy among the suggested Logistic Regression and K-Nearest Neighboring 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 KNN algorithm with p=0.025 (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** | .021 | 0.02 | 2.449 | 18 | .025 | 2.67000 | 1.09036 | .37924 | 4.96076 |
| **Equal Variances not assumed** |  |  | 2.449 | 18.00 | .025 | 2.67000 | 1.09036 | .37924 | 4.96076 |

****

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