Using the Agglomerative Hierarchical Clustering Method to Examine Human Factors in Indonesian Aviation Accidents

Based on the National Transportation Safety Committee (KNKT) Database 1997-2020

Rossi Passarella[1]*, Gulfi Oktariani[2], Dedy Kurniawan[3], Purwita Sari[4]
Department of Computer Engineering, Faculty of Computer Science. Universitas Sriwijaya, Indralaya 30662, Indonesia[1, 2]
Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Palembang 30139, Indonesia[1]
Informatic Management, Faculty of Computer Science. Universitas Sriwijaya, Indralaya 30662, Indonesia[3, 4]

***Abstract*—This study aims to provide a comprehensive source of knowledge regarding aviation accidents in Indonesia caused by human factors, which are the most significant among other causative elements, requiring a detailed assessment of the accident as a result of pilot and co-pilot faults while operating the aircraft. The KNKT website database is still in the form of accident reports. To this end, the retrieved information based on historical data for 23 years of accidents caused by humans by analyzing the data using the clustering approach to gain data insight in the relationship between total flying hours and pilot licenses. The data analysis revealed that, in general, the aircraft operator complied with the CASR standards.**

***Keywords*—Aviation accidents data; pilot's licenses; flying hours; human factor**

I. Introduction

The advancement of technology, particularly in the transportation industry, is usually swift; and one mode of transportation is nearly numbered one in aviation. Due to the relatively rapid time efficiency, security, and safety, aircraft make it easier for individuals to travel between provinces and across nations [1]. However, using this mode of transportation has been associated with a great risk of accidents [2][3]. Several studies on aviation risk have been conducted, especially those related to flight crews [4] [5].

Human casualties, environmental damage, property loss, and psychological impacts are all losses caused by plane crashes. According to Article 3 paragraph (1) of The Minister of Transportation (Indonesia) Regulation No. 77 of 2011, compensation for passengers who perished in aviation accidents amounted to Rp. 1,250,000,000 (one billion two hundred and fifty million rupiahs) per passenger, which is paid to the heirs of the deceased [6]. According to study [3] [7], four elements are found responsible for plane accidents: human factors, environmental factors, facility issues, and engineering considerations. Human factors affecting pilots and co-pilots, aviation security officials, and poor aircraft maintenance staff are among the four. This is followed by environmental conditions such as dense clouds, heavy rain, high winds, and mountains. The properties of the runway and the potential risk of animals being found on it contribute to the facility issues mentioned. Finally, engineering factors are associated with the type of engine utilized as well as aircraft and engine maintenance.

According to study [7], statistics on aviation accidents and incidents in Indonesia gathered from the National Transportation Safety Committee (KNKT) show 26 significant incidents and 15 accidents between 2010 and 2016. There has been a 20% distribution of all events in the last seven years, and several factors were found responsible for such events. The percentage calculation of various causal elements, such as human factors (67.12 percent), technical factors (15.75 percent), environmental factors (12.33 percent), and facilities (4.79 percent) was obtained from the research. Given this context, investigators looked into whether human factors were still relevant after 2016, where to analyze data from the same sources between 2017 and 2020, which revealed that human factors were still the dominant factor in accidents and incidents, accounting for 52.6 percent of all accidents and incidents. This percentage represents a decrease of approximately 14.52 percent from the previous data.

Furthermore, a preliminary experiment was conducted using the same data provided by the KNKT (1997-2020), and using the same approach it was found that the human component still played a role in accidents 23 years ago with a percentage of 46.5 percent [7] [8]. Consequently, if the pilot's performance does not meet the criteria, does not correspond to a standard operating procedure (SOP), or does not correspond to the flying hours, the passenger will be at risk. As a result, the requirements to become a pilot must be understood, of which their license is the most important.

According to [8], the terms of the pilot's license are based on total flight hours under the provisions of the Civil Aviation Safety Regulation (CASR). There are four main licenses in aviation. For a prospective pilot to fly while they are learning, one must receive a Student Pilot License (SPL). When the pilot's flight hours increase, the pilot gets a Private Pilot License (PPL) with a minimum standard flight duration of 50 hours. To earn a Commercial Pilot License (CPL), a minimum standard flight duration of 150 hours must be met, and to obtain an Airline Transport Pilot License (ATPL), a minimum total flight duration requirement of 1500 hours must be met.

*Corresponding Author.

An airline must determine the pilot in charge of flying a commercial aircraft far before the day of departure. What concerns us here is whether accident data from 1997 to 2020 contains improper pilot assignments due to licenses that were inappropriate for flying commercial aircraft flown, which may have led to the accidents. This issue is investigated to uncover the same. To this end, the retrieved information based on historical data for 23 years of accidents caused by human factors by analyzing the KNKT data using the clustering approach to gain data insight based on the variables' total flying hours and pilot licenses. Based on the description above, it is important to categorize total flight hours and pilot licenses such that they may be considered objects with characteristics to analyze.

A structural and hierarchical approach is required to address the research question, categorizing data based on pilot licenses who have encountered accidents or incidents. The agglomerative hierarchical clustering (AHC) method helps generate output in the form of a hierarchical structure that can provide an overview of data comprehensively. AHC is a bottom-up method wherein each data point is initially its own cluster, and as one ascends the hierarchy, additional clusters are amalgamated. In this method, the two neighboring clusters are merged into a single cluster. With this advantage, it allows for the researcher to identify hidden data from clusters so that they may view the distribution of the data and identify whether there are pilots in the KNKT data who have total flight hours.

This research only focuses on archival data of official aircraft accident reports issued by the KNKT Institute and finds out whether the human factor (pilot) shows a correlation with the cause of the accident. While factors outside the discussion are the limitations of this paper.

The structure of the presentation of this research is as follows: Section II describes the materials and methods used; Section III presents the results and discussion; and the conclusions are presented in Section IV.

II. Materials and Method

A. Materials

Researchers used credible and public data sources to obtain information on aviation accidents in Indonesia via the KNKT website database, which is still in the form of accident reports. This collection of accident reports is available on the website, organized by the year of the accident. The report data for each accident includes information on the event's timeline, the pilot's identity, and detailed information regarding the aircraft involved. However, as the data have not been integrated into a unified tabular format, researchers must consolidate accident data into a single database to simplify the analysis.

The researchers only used 341 data points from Indonesian aircraft accidents that occurred between December 19, 1997, and September 15, 2020, accounting for the completeness of each data point. The collected data is then analyzed to produce cluster results and conclusions regarding the link between two factors in an aircraft disaster.

B. Method

The methodology carried out in this research process is described within a research framework, which is illustrated in Fig. 1.

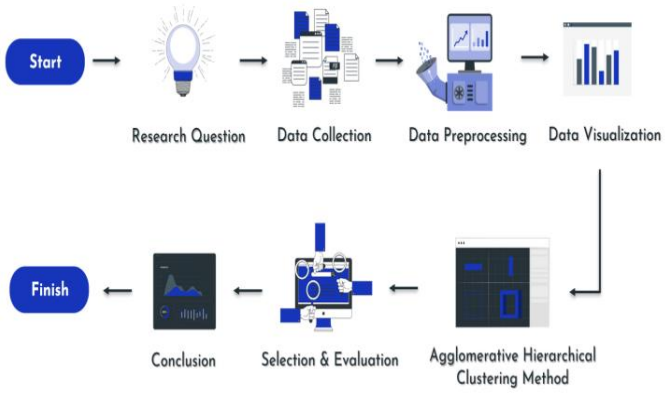

Fig. 1. Research Framework.

The first step of conducting any work of research is to develop a research topic. In making this research inquiry, the goal is to construct a question that is clear, focused, and free of prejudice. This is followed by providing a literature review to comprehend the context of the topic sufficiently. A literature review allows the researcher to construct questions that express the concepts and problems to be explored, also informing the public of the relevance of the outlined issues, which are beneficial to society once addressed, furthering scientific development.

In the second phase of the research, the data used in this study is explained in the previous sub-chapter; data from KNKT about aviation accidents in Indonesia have been used. Data collection of accidents is carried out and then put together into a spreadsheet. Researchers collect essential and relevant information and arrange them in the chronology of accidents. The data collected from accident reports have 16 variables, which include the date, year, province, location, aircraft type, aircraft operator, crew, passengers (PAX), injured, fatalities, type of accident (A/I), flight phase (POA), cause of the accident (PCOA), total pilot's flight hours, type of pilot's license and remarks, altogether culminating in 341 lines of data. After data collection, information is saved with the file extension .csv and imported into the Jupyter notebook program for data analysis.

The third stage is data preparation, which includes multiple processes such as data cleaning, data integration, data transformation, and data reduction. This step involves loading the data into the Jupyter notebook. Data processing is performed, followed by a data cleaning process, which includes the selection of the required variables. Following identification of the required variables, data that includes null or empty values will be cleaned up and outliers in the data will be verified and removed in case they impact the data balance.

The next stage in this study is data transformation. The encoder label is applied in the pilot license variable column to change the data from category to numeric. Furthermore, data reduction will complete the standard scaler procedure at this step, which is required to normalize the value of the variable column of total flight hours and pilot license. Data will be visualized using EDA after completing the data pretreatment stage as part of the pre-modeling phase to display variable data such as total flying hours and pilot licenses, following which, the clustering method is employed for analysis.

The researchers adopted the AHC approach at this stage. This clustering method is divided into several sub-methods for the calculation of proximity between clusters. Researchers only use three methods, which are single-linkage (Closest distance), complete linkage (Farthest distance), and average linkage (Average distance). It then uses Euclidean calculations to compute distance and the silhouette index to estimate the ideal number of clusters. Two variables are employed in this grouping: total flight hours and pilot's license, which will be clustered to determine the link between the two and the frequency of accidents that occur as a result.

Following the above mentioned procedure, a validation test on the number of each cluster found is required to select the best approach. This selection will be based on the value of the validation test index that has a satisfactory performance by analyzing the distribution of the data for each cluster. This is followed by an analysis of research questions based on the relationship between the two variables used, namely total flight hours and pilot license. The method is then evaluated for selection.

The final stage is the conclusion, which will include the results of the AHC method, which involves a validation test value that meets the criteria with the number of clusters used. The results are bound to demonstrate whether or not a relationship relevant to the research question can be established. Furthermore, significant insights will be discovered in the form of information due to the preceding stage's clustering and data distribution.

III. Result and Discussion

A. Data Preparation

Researchers execute variable filters on this preparation data used in this investigation. It is known that there are 16 variables from raw data; however, in this study procedure, only two variables are employed, while the remaining variables are not used. Both variables, total flight hours and pilot's license are utilized.

B. Data Preprocessing

The data on aviation accidents in Indonesia from 1997 to 2020 were collected using the date of the accident reports on KNKT websites. The first step is to erase the missing data values in the variable column used, particularly the total flight hours and pilot's license. After deletion, the total data collected was 177 out of 341 reports. The second procedure is to remove outliers - data with distinctive features that appear different from other data. After removing missing data values, outliers acquired 156 data points from a total of 177 data points.

TABLE I. Data based on Pilot's License

No.	Pilot Licenses	Numbers of Data
1	ATPL	99
2	CPL	45
3	SPL	8
4	PPL	4

Data that has already been cleaned will be transformed into a suitable form for processing. The first step is to examine the various types of pilot licenses including ATPL, CPL, SPL, and PPL. The results are shown in Table I, with data from four pilot licenses.

This step is followed by the creation of an encoder label for the type of pilot license. ATPL (0), CPL (1), PPL (2), and SPL (3) are the designations for each type of license. In this study, data reduction is used to reduce dimensions in data so that later on, data will be comprehensive; this would also retain the integrity of the data as much as possible. This reduction data strategy is a standard scaler that seeks to normalize the data so that massive variations may be avoided. It is used in this study to standardize the variables - total flight hours and pilot's license value.

C. Descriptive Statistics

Following the completion of the data preparation stage, descriptive statistics of the data were used as an additional tool within the data processing stage before proceeding to the data visualization stage. This study employs two types of variables: quantitative and qualitative data variables. The quantitative data variables in this study are variables with numerical numbers - the total flight hours of pilots. Qualitative data variables are categorical variables - pilot licenses. The results are produced using descriptive statistics based on the quantitative and qualitative data factors, as shown in Table II.

According to Table II, the mean or average value of the accurate data for the quantitative data variable (total pilot flight hours) acquired is 7,024.58 hours. Furthermore, for the dispersion value (data spread to the average value) acquired by 0.69, this variable's median value was found to be 5,935 hours. This signifies that the variable corresponds with homogenous data with a tiny dispersion value - a minimum value of 20 hours and a maximum value of 17,547 hours. It also implies that the missing value for this variable is already worth 0, which means that there is no lost or empty data. While the type of qualitative data variable (pilot license) is categorical, it is converted to numerical using the encoder label. The pilot license data has a mean value of 0.50.

Furthermore, the median value of this variable is 0, indicating that ATPL is the midpoint value of the pilot license variable. Aside from the dispersion value, the average value was found to be 1.59, indicating that this variable corresponds to heterogeneous data with a high dispersion value, a minimum value of 0, and a maximum value of 3. For this variable, a missing value is already worth 0, indicating that no data was lost or is empty.

TABLE II. DESCRIPTIVE STATISTICS OF DATA

Variable Types	Mean	Median	Dispersion	Min	Max	Missing Value
Quantitative	7024.58	5935	0.69	20	17547	0(0%)
Qualitative	0.50	0	1.59	0	3	0(0%)

D. Data Visualization (EDA)

Following the data preprocessing and descriptive statistics phases, the EDA step is used to comprehend the contents and components of the data. The EDA approach employed in this study is univariate and bivariate; univariate analysis is used to distribute data from a single variable while bivariate analysis demonstrates the link between two variables [9]. To comprehend data trends, both analyses are used to examine the distribution of data for varying total flight hours and pilot licenses in detail. As follows, pie charts are used for univariate analysis and scatter plots are used for bivariate analysis.

The Pie Chart is used to determine how much category distributes on a pilot's license. According to Fig. 2, the most significant proportion of ATPL (0) license types was 63.2 percent, followed by 29.0 percent for the CPL (1) category, 2.6 percent for the PPL (2) category, and 5.2 percent for the SPL (3) type. The scatter plot is used to illustrate data distribution, to examine how data spread links between the variable total flight hours and the pilot's license.

According to Fig. 3, it may be noted that the patterns across the two variables do not have a link as no tendency of specific values can be observed in common. Each pilot's license has its unique pattern that appears to be completely unrelated to the value of each pilot's license in the crash data.

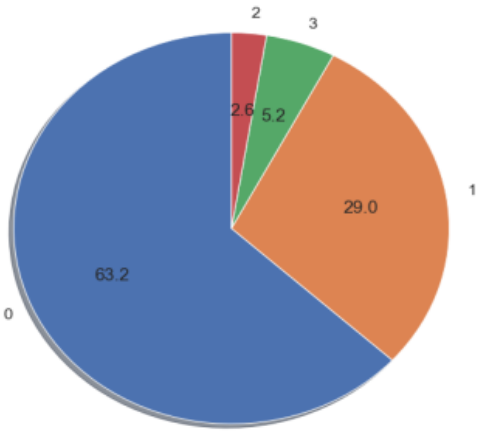

Fig. 2. Pilot License Data Visualization.

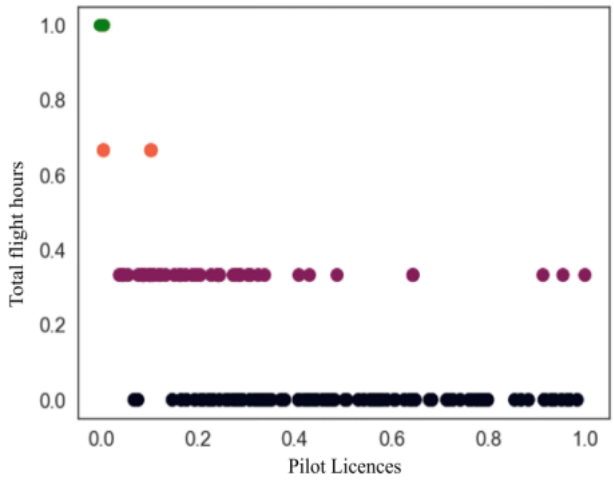

Fig. 3. Visualization of Total Flight Hours based on Pilot's License.

E. Cluster Test

At this juncture, it is important to mention that the test used three predefined agglomerative strategies. First, the distance is calculated using the Euclidean formula, then the ideal number of clusters is estimated using the silhouette index technique, the amount of data spread for each cluster is examined, and finally, the dendrogram of each approach is visualized.

The following are the findings of the validation test of the number of clusters using the silhouette index technique with a range of cluster values 2-4. Using two clusters yielded an average value of 0.5156, three clusters yielded an average value of 0.5004, and four clusters yielded an average value of 0.5500. Based on the computation of the silhouette index, if the average value is close to the value of one, the data grouping is considered better [10].

F. Hierarchical Agglomerative Method Test

The grouping approach employs the three techniques selected to acquire the data dispersed across clusters and provide dendrogram visualizations based on the number of clusters formed from each AHC method.

1) Single linkage: The single linkage approach is used to group two variables with a distance between cluster pairs determined by the two nearest objects between separate clusters [11]. According to Table III, each cluster contains a varied quantity of data, and the data in each cluster is far apart and unbalanced. Cluster 3 has the least data with 4 data points, while Cluster 2 has the most data with 99 data points. The dendrogram construction employs a single linking approach with four clusters connected by a pair of points with the closest distance between each other, as depicted in Fig. 4. (a).

TABLE III. NUMBER OF EACH SINGLE LINKAGE CLUSTER

Cluster	Numbers of Data
1	45
2	99
3	4
4	8

2) Complete linkage: The distance between clusters has been computed on the furthest distance between a pair of items [12] [13]; the quantity of data per cluster and visualization results (using a dendrogram) was determined with the Euclidean formulae for four clusters. The data of each cluster corresponded with a ratio that was less balanced than the single linkage technique due to the quantity of data used in the complete linkage method (See Table IV). Cluster 4 has the least quantity of data (12) while Cluster 1 has the most data (72).

TABLE IV. NUMBER OF EACH COMPLETE LINKAGE CLUSTER

Cluster	Numbers of Data
1	72
2	42
3	30
4	12

The output of the dendrogram implementation employs the entire linkage approach, with four clusters connected from the longest distance from a pair of objects seen in Fig. 4. (b).

3) Average linkage: The average linkage computes the average distance between all pairs of data points from various clusters [14]. The amount of data per cluster and the visualization results used for four clusters were determined using Euclidean equations. The number of clusters in this technique is not found to be highly significant, as seen in Table V. In comparison to the whole linkage and single linkage approaches, the amount of data in each cluster is more evenly distributed. Cluster 1 has the smallest amount of data, whereas Cluster 2 has the largest. The average linkage approach is used to achieve the results of the dendrogram implementation, as shown in Fig.4(c).

TABLE V. NUMBER OF EACH AVERAGE LINKAGE CLUSTER

Cluster	Numbers of Data
1	30
2	53
3	42
4	31

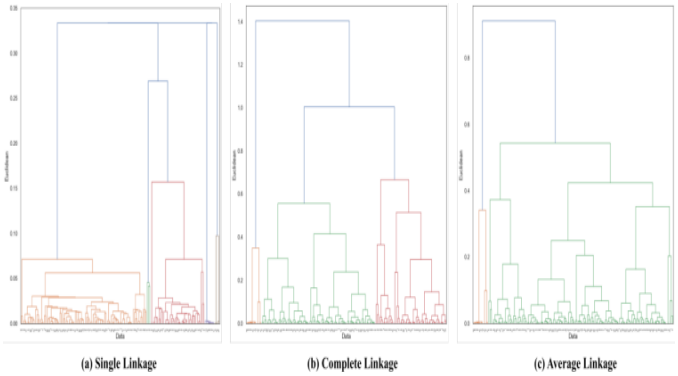

Fig. 4. Visualization Dendrogram Method: (a) Single-linkage (b) Complete Linkage (c) and Average Linkage.

G. Index Validation Test

The validation test of the score value in clustering is used to determine the optimal AHC approach. The silhouette index matrix was used in the first measuring procedure, the Davies Bouldin index in the second, and the Calinski Harabasz index in the third.

The data grouping is considered better in silhouette index calculations if the index value is close to one [15]. The results of calculations using the silhouette index based on Table VI show that the index values for the three grouping techniques differ. Compared to the single and complete linkage, the average linkage technique yields the highest index value with 0.5623 or 56% being calculated. The silhouette index method was then used to obtain the best method on the validation test, namely average linkage.

The Davies Bouldin index is used for the second measurement. According to [16], the index value approaches zero, implying that the data grouping in one cluster is better. The Davies Bouldin index calculations in Table VI show that the three techniques have varying index values. The average linkage technique has the lowest index value of 0.5644 or 56%. As a result of the validation test using the Davies Bouldin index, the best technique obtained is average linkage.

The Calinski Harabasz index is used in the third measurement, wherein the index value is not limited. The higher the index value, the better the data grouping in one cluster [17]. Table VI shows the results of the Calinski Harabasz index calculation. It is well known that the index values of the three techniques differ. The average linkage technique produces the highest index value of 284.68, while the single linkage technique produces the lowest index value of 93.37. Using the Calinski Harabasz index, the best technique is determined using the validation test, which was found to be average linkage. For the three validation test methods, all the measurements showed elicited 100 percent, indicating that average linkage is the best method for viewing data insights into licenses and total flight hours.

TABLE VI. INDEX VALIDATION TEST RESULTS

Methods	Single	Complete	Average
Silhouette	0.4838	0.4045	0.5623
Davies Bouldin	0.6238	0.7421	0.5644
Calinski Harabasz	93.37	159.09	284.68

H. Analysis of Cluster Average Linkage Results

This study shows the results of each cluster in relation to the total flight hours and pilot licenses based on the type of the pilot license used - ATPL, CPL, SPL, and PPL, which were mapped based on the number of clusters. The average value has also been based on the number of clusters. Table VII displays the number of pilot license categories in each cluster and the total number of pilots' flight hours.

After selecting the agglomerative method approach with average linkage, the researcher noticed two accidents in the fourth cluster involving pilot licenses under the ATPL category (Table VII). While the pilot involved in the fourth cluster had

an average flight time of 1071.67 hours, which when combined should have been more than 1071.67 hours, the ATPL standard is at least 1500 hours. Such anomalies much be investigated as the total rata-average of these four licensing classes is 1071.67 hours, while certain anomalies may be assumed to have a duration of over 1071.67 hours. Thus, a deeper delve into Cluster 4 was necessary. Based on the established statistics, researchers found two ATPL-type permits with a total flight hour of fewer than 1500 hours. Despite CASR standards requiring ATPL type pilots to have a minimum flight hour requirement of 1500 hours, two pilots in Cluster 4 did not meet the minimum flying hours required by their pilot's license, with total flight hours of 1200 hours and 1339 hours, for each pilot. The Cluster 4 results are shown in Fig. 5, illustrating the ATPL (0) licenses that were found to be different from the rest of the data.

TABLE VII. NUMBER OF PILOT'S LICENSES AND AVERAGE TOTAL PILOT HOURS

Cluster	Pilot Licenses				Average of Total Data Pilot's Flight (Hours)
	ATPL (0)	*CPL (1)*	*PPL (2)*	*SPL (3)*	
1	27	3	0	0	14625,23
2	33	20	0	0	4634,84
3	37	5	0	0	9060,54
4	2	17	4	8	1071,67

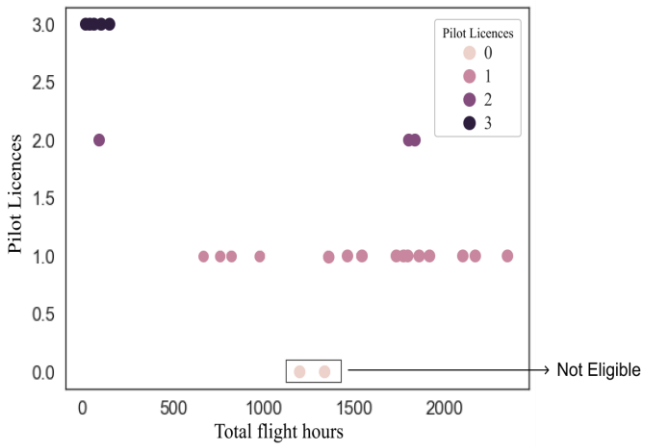

Fig. 5. Visualized Data Cluster 4 Average Linkage.

IV. CONCLUSION

This study was carried out to provide a comprehensive source of knowledge regarding aviation accidents in Indonesia caused by human factors, which are the most significant among other causative elements, requiring a detailed assessment of the accident as a result of pilot and co-pilot faults while operating the aircraft. The approach used to analyze KNKT data was clustering algorithms to identify insightful data in the relationship between total flying hours and pilot's license, based on the historical data of accidents caused by human factors over 23 years.

The data analysis revealed that, in general, the aircraft operator complied with the CASR standards. However, using clustering, it was discovered that the pilot had a license that did not match with the flown aircraft in 1.3 percent of the 156 accidents. Another finding revealed that, in comparison with other causes, human-caused accidents (pilots) had the highest proportion - with 46.5 percent from 1997 to 2020. Meanwhile, the level of representation of the analyzed data was only 46 percent as only 156 out of 341 accidents over 23 years were processed due to the relevant observation values on license data variables and flight hours.

ACKNOWLEDGMENT

We thank those who have helped us complete this study.

REFERENCES

[1] R. Passarella, and S. Nurmaini, 2022 "Behavioral Evidence of Public Aircraft with Historical Data: The Case of Boeing 737 MAX 8 PK-LQP," *Journal of Applied Engineering Science*, vol.20, no.4.

[2] Y. Wei, H. Xu, Y. Xue, and X. Duan 2020 "Quantitative assessment and visualization of flight risk induced by coupled multi-factor under icing conditions," *Chinese J. Aeronaut.*, vol. 33, no. 8, pp. 2146–2161. DOI: 10.1016/j.cja.2020.03.025.

[3] R. Passarella., and S. Nurmaini 2022 "Data Analysis Investigation: Papua is The Most Unsafe Province in Indonesia for Aviation: An Exploratory Data Analysis Study from KNKT-Database Accidents and Incidents (1988-2021)," *Journal of Engineering Science and Technology Review (JESTR)*, vol. 15, no.3, pp 158-164. DOI: 10.25103/jestr.153.17.

[4] S. Gentile, A. Furia, and F Strollo, 2020 "Aircraft pilot licence and diabetes," *Diabetes Research and Clinical Practice*, vol 161, DOI: 10.1016/j.diabres.2020.108047.

[5] M. Efthymiou, S. Whiston, JF. O'Connell, and GD. Brown, 2021 " Flight crew evaluation of the flight time limitations regulation," *Case Studies on Transport Policy*, vol 9, no. 1, pp 280-290, DOI: 10.1016/j.cstp.2021.01.002.

[6] R. Kristiawan, Rolan, Abdullah 2018, "Faktor penyebab terjadinya kecelakaan kerja pada area penambangan batu kapur unit alat berat pt. semen padang.," *J. Bina Tambang*, vol. 5, no. 2, pp. 11–21.

[7] Rahimudin 2015, "Analisis Faktor-Faktor Penyebab Kecelakaan Pesawat Udara Komersil Di Indonesia Pada Tahun 2002 Sampai Dengan Tahun 2012," vol. 8, pp. 82–83.

[8] Y. Xue and G. Fu 2018, "A modified accident analysis and investigation model for the general aviation industry: emphasizing on human and organizational factors," *J. Safety Res.*, vol. 67, pp. 1–15. DOI: 10.1016/j.jsr.2018.09.008.

[9] D. C.Hoaglin 2015, "Exploratory Data Analysis: Univariate Methods," *Int. Encycl. Soc. Behav. Sci. (Second Ed.*, pp. 600–604. [Online]. Available: DOI:10.1016/B978-0-08-097086-8.42125-2.

[10] K. Gonzalez, and S. Misra, 2022 "Unsupervised learning monitors the carbon-dioxide plume in the subsurface carbon storage reservoir,"Expert Systems with Applications, vol. 201. DOI:10.1016/j.eswa.2022.117216.

[11] F. Ros, and S. Guillaume,2019 "A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise," *Expert Systems with Applications*, vol. 128, pp 96-108. DOI:10.1016/j.eswa.2019.03.031.

[12] D. Krznaric, and C. Levcopoulos, 2002 "Optimal algorithms for complete linkage clustering in d dimensions," Theoretical Computer Science, vol. 286, no.1, pp 139-149. DOI:10.1016/S0304-3975(01)00239-0.

[13] P. Govender and V. Sivakumar 2020, Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019), vol. 11, no. 1. Turkish National Committee for Air Pollution Research and Control.

[14] L. Ramos Emmendorfer and A. M. de Paula Canuto 2021, “A generalized average linkage criterion for Hierarchical Agglomerative Clustering,” *Appl. Soft Comput.*, vol. 100, p. 106990. DOI: 10.1016/j.asoc.2020.106990.

[15] M. Gagolewski, M. Bartoszuk and A. Cena, 2021 "Are Cluster validity measure (in)valid ?", *Information Sciences*, Vol. 581,pp 620-636.

[16] DL. Davies and DW. Bouldin, 1979 "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, No.2.

[17] J. Baarsch and M.E. Celebi, 2012 " Investigation of Internal Validity Measures for K-Means Clustering". *Proceedings of the International MultiConference of Engineers and Computer Scienctists* (IMECS), Vol.1. March 14-16, Hong Kong.