STRAHOS 2022
19th Seminar of Track Management
13 and 14 October 2022, Poprad, Slovakia

Special Session 'New Railway Structures'

THE ANALYSIS OF FACTORS AFFECTING THE OCCURRENCE OF INCIDENTS WITH RAIL VEHICLES IN 2007-2021 IN POLAND

Artur Budzyński[1], Aleksander Sładkowski

1. Introduction

Traffic accidents are undesirable. The problem affects the whole world. Accidents are the cause of death and injury to many people. Road accidents are the result of a combination of factors. There are many of these factors. The causes of accidents can be divided into direct and indirect ones. There are human, technical, and external factors. The human factor is making a mistake. Indirect factors influence human mistakes. Fatigue is an indirect cause of the human error. Lack of fatigue can be a factor that can prevent an accident from an external factor or a technical one.

In [1] research was done on people from South Wales. The survey was used with the use of questionnaires. The dependence of participation in the accident on the consumption of caffeine was checked. The accident rate among caffeine-consuming people was 1.9%, and 3.3% in the non-caffeine-consuming group. More likely to be involved in road accidents are younger people, those in poor health, stressed out, and taking greater risks.

Accidents in Sudan were analyzed in [2]. The data was from questionnaires made by an accident victim and eyewitness. Police reports were also used. The accident was affected geographically. Most of it happened in the capital of the country - Khartoum 46.7%. The influencing factor was the scene of the incident. Accidents at crossings accounted for 46.8% of all. The causes of the accidents were driver errors. One of the mistakes is overspeeding. This is a deliberate error that sets it apart from the rest, accounting for 14.6% of accidents.

The article [3] deals with the subject of methodology in predicting accidents, focusing on psychology. According to the authors, few studies understand the methodological complexity of predicting accidents. Improving road safety still requires research and work.

[1] Silesian University of Technology. Faculty of Transport and Aviation Engineering. Department of Rail Transport. Krasiński 8, 40-019 Katowice, Poland

Fig. 1 shows the cumulative line graph of the number of accidents in individual countries. European countries and the Channel Tunnel separately are presented. The number of accidents is visibly decreasing from year to year, which shows that safety is improving. The trend in individual years is similar for all countries. For example, in 2014, the number of accidents increased in all countries compared to the previous year. The decreasing number of accidents is a socially and economically beneficial phenomenon.

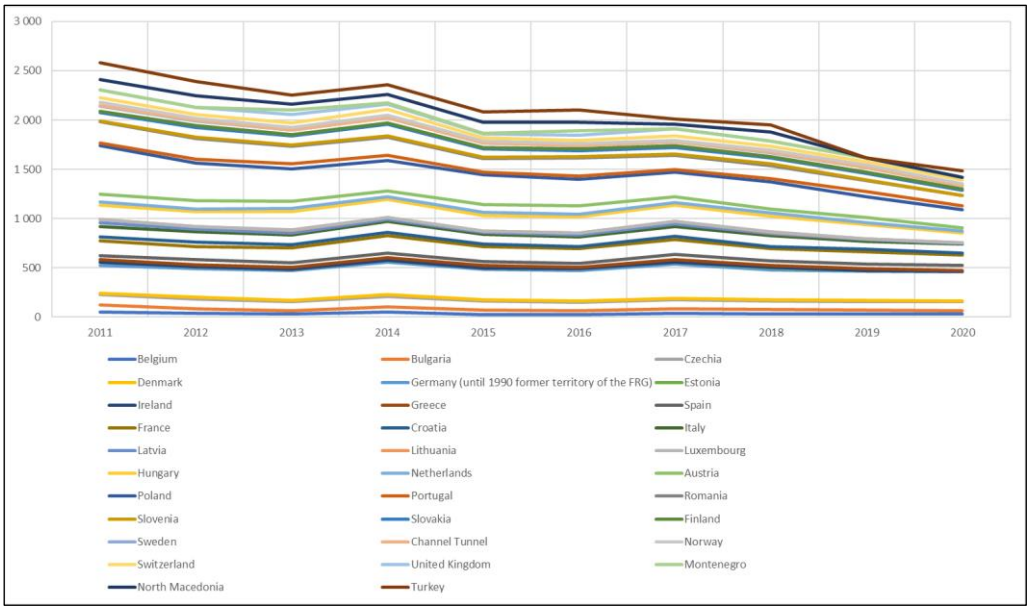

Fig. 1 – Graph of rail accidents by country and by year [4]

2. Methodology

The project is written using the Python [5] programming language. The code is written using Jupyter Notebook [6]. It is an interactive web-based environment for writing and sharing code. The Pandas [7] library was used for data processing. The creation of new features was done by summing the column values in the DataFrame Pandas. MatplotLib [8] and Seaborn [9] libraries were used for data visualization. MatplotLib was used to create bar and line charts. Seaborn was used to create a heatmap. The data used for the study concern accidents that occurred on the territory of Poland from 2007 to 2021. The data has been downloaded from the website [10] based on the data of the General Police Headquarters. The data concerns accidents of rail vehicles with other road users. The Polish police distinguish between 15 types of vehicles involved in the incident. This study focused on two types of rails: "Train" and "Tram, trolleybus' '. A trolleybus is not a rail vehicle; however, the police do not distinguish between trolleybuses and trams as separate types.

Referring to the defined term "injured" of the Central Statistical Office in Poland [11]. A person injured is a person who was not killed on the spot due to an accident or died within 30 days, but sustained injuries that usually require treatment. A person with a hard injury is injured who was hospitalized for more than 24 hours as a result of an accident. Slightly injured is a person injured, excluding the dead and seriously injured. People who are less injured, such as minor cuts or bruises, are usually not recorded as injured. Based on this information, there is a fine line between classifying a person as seriously injured or dead. For example, a person who dies on the 31st day after the accident is statistically seriously injured. There is a fine line between classifying a person

as seriously injured or slightly injured. After the accident, the participants of the accident are accompanied by many emotions related to stress. A person who feels good immediately after the event and does not require medical attention may feel worse when the emotions stop working. For the purposes of the project, it was decided to calculate an additional variable "victims" as the sum of the dead, hard injured, and slightly injured.

3. Results and discussions

Fig. 2 shows the annual variability of incidents with trains. The victims are the sum of the slightly injured, hard injured, and the dead. Year-by-year mean percent change was calculated for all variables. The average percentage trend for all variables is minus. The number of incidents decreased on average by 4.36% per year. The total number of victims decreased on average by 5.88% per year.

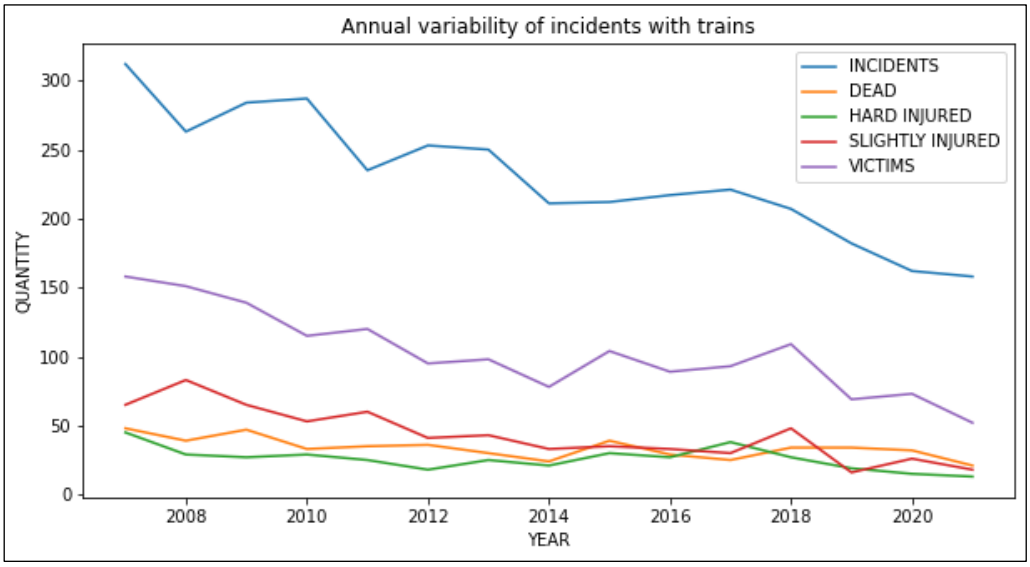

Fig. 2 – Annual variability of incidents with trains, own study based on [8]

Among them, death decreased on average by 2.64% per year. Hard injuries decreased on average by 5.89% per year. Slightly injured on average of 2.35% per year. The number of incidents involving trains decreased in the years 2007-2021, which is an undisputedly favorable phenomenon.

Fig. 3 shows the annual variability of incidents with trams/trolleybuses. Year-by-year mean percent change was calculated for all variables. The average percentage trend for all variables is minus. The number of incidents decreased on average by 4.34% per year. The total number of victims decreased on average by 8.65% per year. Among them, death decreased on average by 6.90% per year. Hard injuries decreased on average by 3.94% per year. Slightly injured on average of 9.29% per year. The number of incidents involving tram/trolley buses decreased in the years 2007-2021, which is an undisputedly favorable phenomenon. The overall number of incidents is declining at a similar rate in both cases. The number of deaths as a result of accidents involving trams/trolleybuses drops faster than in cases involving a train. The number of deaths as a result of accidents involving trams/trolleybuses drops more than 2 times faster than in cases involving a train. The number of hard injuries as a result of accidents involving a train drops faster than in cases involving trams/trolleybuses. The number of slightly injured as a result of

accidents involving trams/trolleybuses drops more than 3 times faster than in cases involving a train.

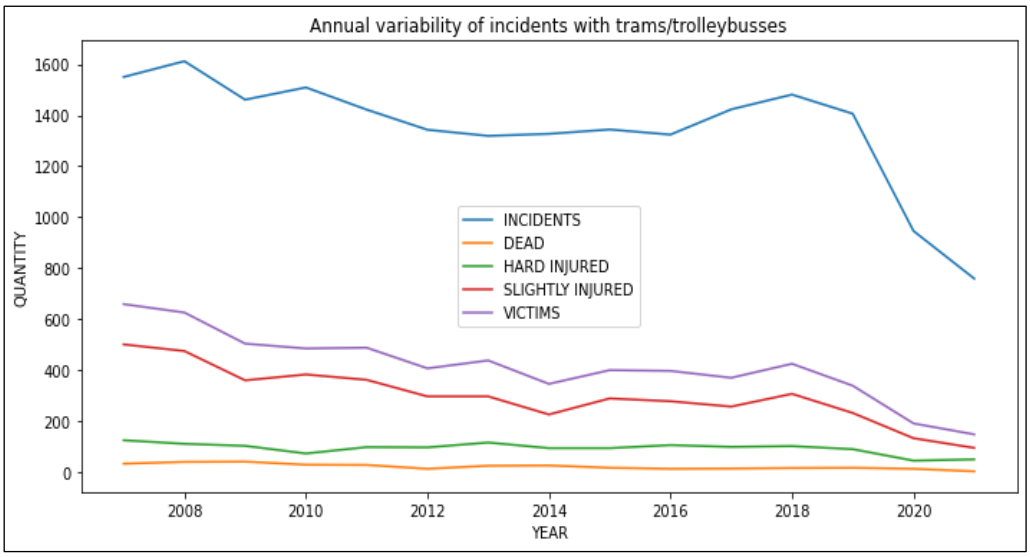

Fig. 3 – Annual variability of incidents with tram/trolley buses, own study based on [8]

Fig. 4 shows correlation matrix dependencies on accidents with trains. The purpose of creating this was to investigate the relationship between individual variables. There is a minus high correlation between the year and the rest of the variables. Based on the above information it is clear. The number of incidents correlates more with the total number of victims than with individuals. The use of the sum of victims in statistical analysis has proven beneficial. There is less correlation between the borderline descriptions of the victims. The correlation between light injured and dead is 0.68. The correlation between light injured and hard injured is 0.52. The correlation between death and hard injury is 0.48.

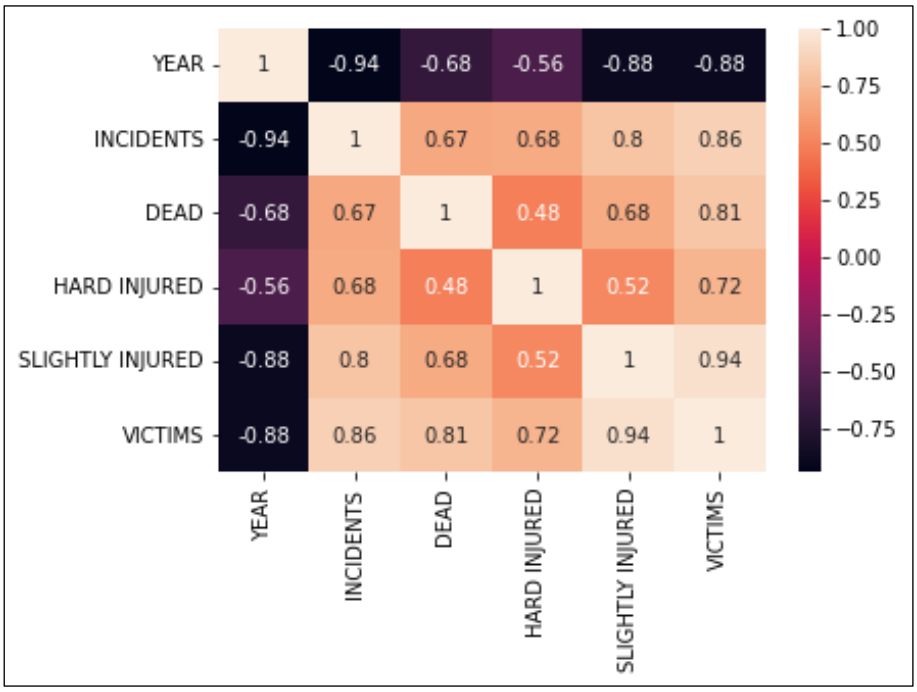

Fig. 4 – Correlation matrix of variables related to incidents with trains, own study based on [8]

Fig. 5 shows correlation matrix dependencies on accidents with trams/trolleybuses. The number of incidents correlates more with the total number of victims than with individuals. There is less correlation between the borderline descriptions of the victims. The correlation between light injured and dead is 0.8. The correlation between slightly injured and hard injured is 0.74. The correlation of death with hard injury is 0.54. Summing up, the number of incidents with victims correlates more in the case of accidents with trains than with trams/trolleybuses. Train accidents are more varied.

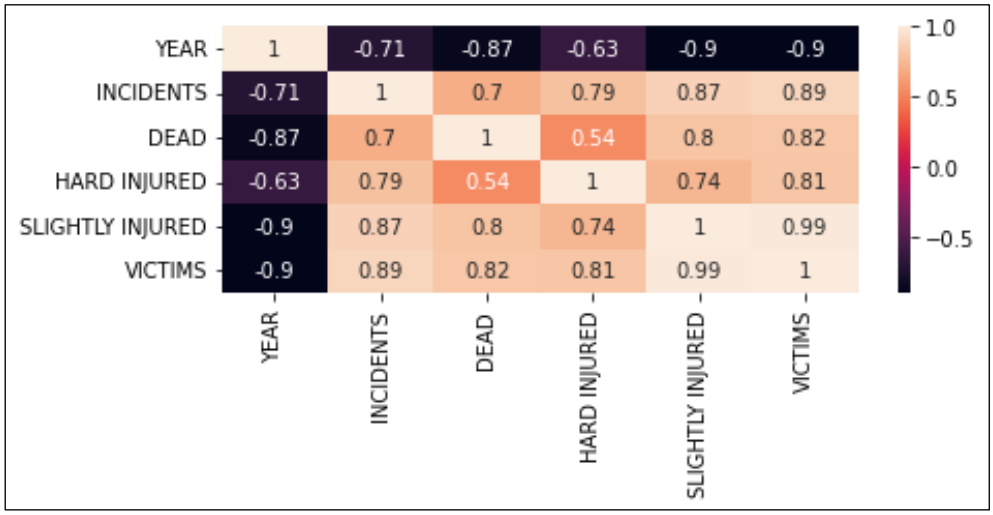

Fig. 5 – Correlation matrix of variables related to incidents with trams/trolleybuses, own study based on [8]

Fig. 6 shows bar chart quantity incidents with trains depending on the month. The chart has 4 months and there is a trend change. These are January, March, July and November. The months in which the trend changes from rising to falling are July, where the temperature is highest, and January, where the temperature is the lowest. The highest number of accidents is in July 254. The lowest number of accidents was in March 228. In Poland, July has the highest temperatures of the year. January is the lowest temperature of the year. On the basis of this, it is worth considering in detail the influence of temperature on the occurrence of an accident. Dependency incidents with trains depending on weekdays. Standard deviation was 41.77.

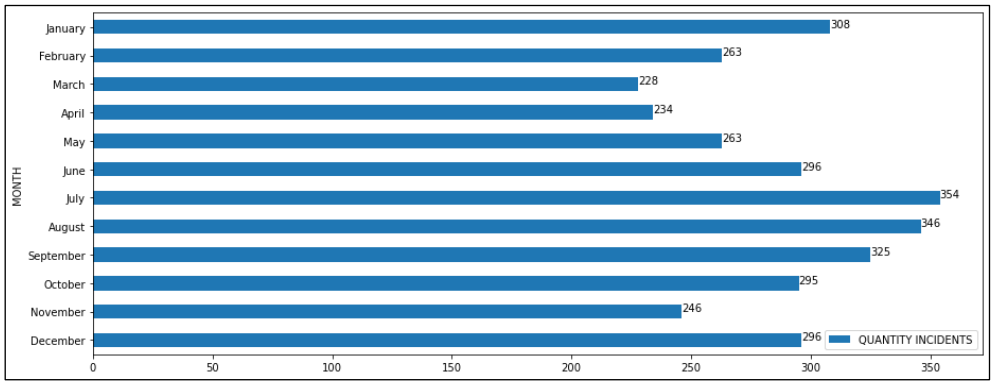

Fig. 6 – Incidents with trains/trolleybuses depending from month, own study based on [8]

Fig. 7 shots bar chart quantity incidents with trams/trolleybuses depending from month. Accidents involving trams/trolleybuses are not temperature dependent like trains. Most incidents involving trams/trolleybuses were in October 2137. The lowest number of incidents involving trams / trolleybuses was in July: 1,310. Unlike in the case of trains. The standard deviation of incidents involving trams/trolleybuses was 224.42.

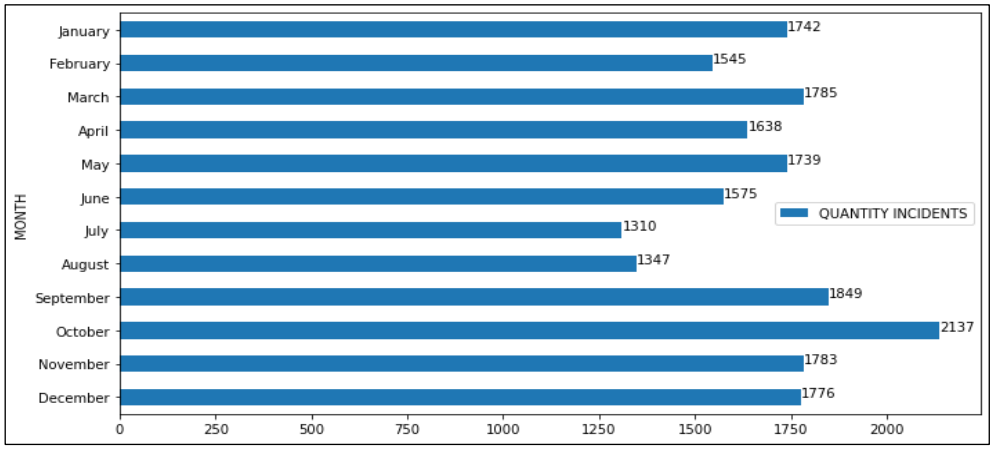

Fig. 7 - Incidents with trams/trolleybuses depending from month, own study based on [8]

Fig. 8 shows bar chart quantity incidents with trains depending on weekdays. *Fig. 9* shows bar chart quantity incidents with trams/trolleybuses depending from weekday. On Sunday, there were the fewest incidents in both cases. Saturdays were more than Sunday but less than Monday to Friday. This may be because days from Monday to Friday are working days when many plants and institutions are working. Saturdays and Sundays are often days off. This is important information for modeling. It is proposed to distinguish 3 categories in terms of the week: working day, Saturday, Sunday.

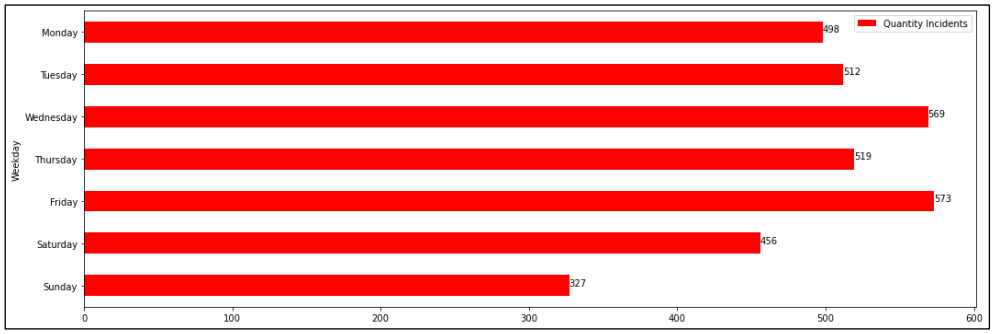

Fig. 8 – Incidents with trains depending from weekday, own study based on [8]

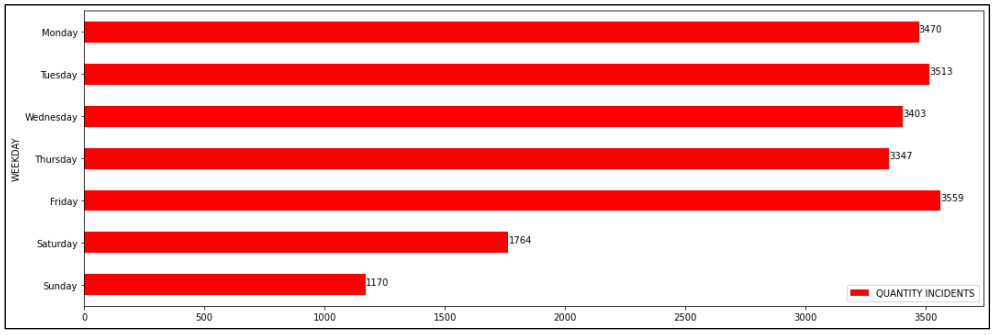

Fig. 9 – Incidents with trams/trolleybuses depending from weekday, own study based on [8]

Fig. 10 shows line chart quantity incidents with trains depending on time. *Fig. 11* shows bar chart quantity incidents with trams/trolleybuses depending from weekday. The number of incidents can be related to the amount of traffic. There are fewer incidents at night than during the day. The dynamic increase in the number of incidents takes place after 4:00 a.m. This is related to the commute that starts at 6:00 AM from 4:00 PM. there is a gradual decrease in the number of incidents. The course of the train incidents chart is more irregular than tram/trolleybuses, because train incidents are less. The above information is important for model training. It is proposed to assign an hour to the accident. It can create a new "time of day" feature and assign incidents a value like "night", "morning" and "afternoon".

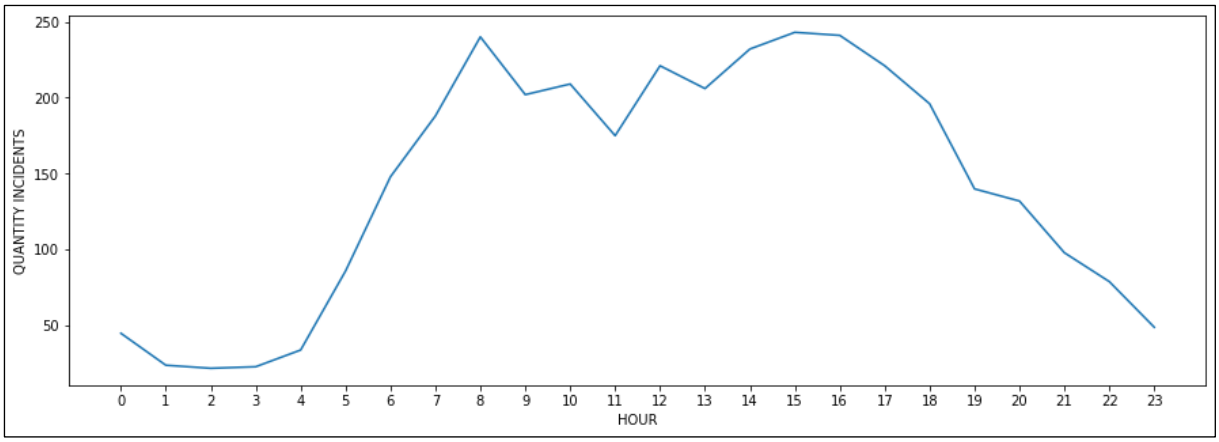

Fig. 10 – Incidents with trains depending on time, own study based on [8]

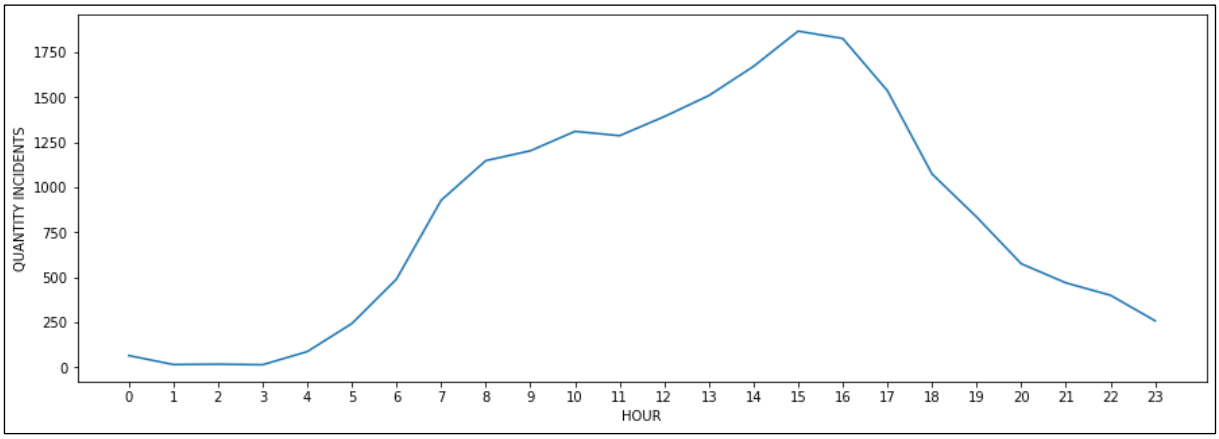

Fig. 11 – Incidents with trams/trolleybuses depending on time, own study based on [8]

4. Machine Learning Prediction Model

It is proposed to train a classification machine learning model. The model is to distinguish between accident and collision events. An experiment was conducted to compare 4 machine learning models. The results are shown in *Fig. 12*. The differences were slight. The feature engineering process was the key to the result. The factorize function of data processing for use in the model was used here. The best model is GradientBoostingClassifier with accuracy of approximately 0.939. The top 20 features for the GradientBoostingClassifier model are shown in *Fig. 13*. The type of incident is a key feature that influences an accident. Different types cause different harm. The Characteristics of the place of the event works similarly. Features: Voivodeship, GPS x,

County, Road Number, Street, Intersection determine the dependence of the accident on the place of its occurrence. GPS x determines the latitude that is influenced. GPS y is longitude. Dayofyear, Year, and Weekday specify the date. Type of road, the condition of the surface, KM HM (Chainage), Speed limit are determined by the conditions at the scene of the accident. The accident depends on variables of various types: date, location and conditions. Data for this project is available in open access [12].

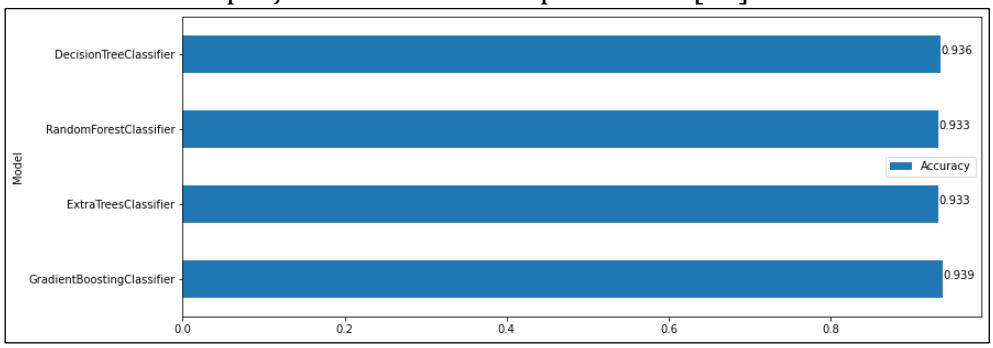

Fig. 12 – Comparison of 4 models according to accuracy

Weight	Feature
0.6959 ± 0.5380	The type of incident
0.0744 ± 0.2770	Characteristics of the place of the event
0.0318 ± 0.2628	Voivodeship
0.0295 ± 0.0881	GPS x
0.0199 ± 0.2802	County
0.0196 ± 0.1499	Other reasons
0.0179 ± 0.1121	Year
0.0167 ± 0.1207	Road Number
0.0144 ± 0.0362	Street
0.0130 ± 0.0426	Intersection
0.0106 ± 0.0850	Community
0.0098 ± 0.1157	Dayofyear
0.0069 ± 0.0534	Speed limit
0.0059 ± 0.0528	Area
0.0057 ± 0.0386	KM HM (Chainage)
0.0052 ± 0.0593	Type of road
0.0050 ± 0.0387	The condition of the surface
0.0040 ± 0.0545	GPS y
0.0025 ± 0.0436	Weekday
0.0024 ± 0.0628	Marking
	… 12 more …

Fig. 13 – Top 20 features for the GradientBoostingClassifier model

5. Conclusions

The division of trams and trolleybuses into two separate categories will allow for more detailed analyzes of the Polish police. It is worth analyzing the overall victim's ratio as the sum of the dead, seriously injured, and slightly injured. The use of universal accident data recording for police forces around the world will allow better comparisons of data between countries. The solution, which is easier to implement, is the use of

a universal method of recording accident data for the police in the European Union. It is recommended that data be made open access by police forces around the world to consolidate research efforts. The number of accidents drops from year to year. Modeling an accident allows one to better understand the dependencies that affect its occurrence. Accidents are absolutely negative, and efforts to limit them and reduce any negative impacts are valuable; the accident depends on variables of various types: date, location, and conditions. It is possible to predict whether a given incident will be a collision or an accident. The best model is GradientBoostingClassifier with the accuracy of approximately 0.939. Data for this project is available in open access.

The methodology of recording accident data used by the Polish Police was analyzed. The European Union is an area that integrates many areas. Accident data collection could be one of these areas. In subsequent studies, it is worth thoroughly analyzing the system in which data on accidents in all European Union countries are recorded. This will allow conclusions to be drawn about good and bad practices. Such action will allow the commencement of discussions on the development of a common accident data recording system for the European Union. Saving data, in the same way, will allow researchers to efficiently model and analyze data for the entire area.

Acknowledgments

The project was created thanks to cooperation with the Polish Police Headquarters. The police made the accident data available for investigation. Publication supported by the Excellence Initiative Research University programme of the Silesian University of Technology, 2021.

This work was supported by the Grant No. 22120015. The project is co-financed by the Governments of Czechia, Hungary, Poland and Slovakia through Visegrad Grants from International Visegrad Fund. The mission of the fund is to advance ideas for sustainable regional cooperation in Central Europe.

References

[1] SMITH, A., Caffeine and road traffic accidents. World Journal of Pharmaceutical and Medical Research. 2021. Vol. 7. No. 12, P. 1-4.

[2] HASSAN, A.A., OSMAN, A.H., ALHASSAN, A., HUSSAIN, G., AHMED, A-H., OSMAN, M., AHMED, H., ELHASSAN, A., Epidemiology of road traffic accidents in Sudan. European Academic Research. 2016. Vol. 3. No. 11. P. 12285-12299.

[3] WÅHLBERG, A.E., Some methodological deficiencies in studies on traffic accident predictors. Accident Analysis & Prevention. 2003. Vol. 35. No. 4, P. 473-486.

[4] Rail accidents [online]. [cit. 2022-068-31]. Available on internet: https://ec.europa.eu/eurostat/en/

[5] VAN ROSSUM G., DRAKE, F.L., *Python 3 Reference Manual*, 2009. ISBN 9781441412690.

[6] KLUYVER, T., RAGAN-KELLEY, B., PEREZ, F., GRANGER, B., BUISSONNIER, M., FREDERIC, J., KELLEY, K., HAMRICK, J., GROUT, J., CORLAY, S., IVANOV, P., AVILA, D. ABDALLA, S. Jupyter notebooks – a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas. 2016. P. 87-90.

[7] MCKINNEY, W., Data structures for statistical computing in Python. In: Proc. of the 9th Python in Science Conf. 2010. P. 56-61.

[8] HUNTER, J. D., Matplotlib: A 2d graphics environment. Computing in Science & Engineering. 2007. Vol. 9. No. 3. P. 90-95.
[9] WASKOM, M.L., Seaborn: statistical data visualization. Journal of Open Source Software. 2021. Vol. 6(60). No. 3021. P. 1-4.
[10] Accident and collision record system [online]. [cit. 2022-06-09]. Available on internet: http://sewik.pl/
[11] The term "injured" used in official statistics according to the central statistical office in Poland [online]. [cit. 2022-06-09]. Available on internet: https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/3829,pojecie.html?pdf=1
[12] Machine learning model for classification of accidents – GitHub Repository [online]. [cit. 2022-06-09]. Available on internet: https://github.com/BudzynskiA/RA