Unsupervised Machine Learning for Pattern Identification in Occupational Accidents

Fatemeh Davoudi Kakhki[1,3], *Steven A. Freeman*[2], *Gretchen A. Mosher*[2]

[1] Department of Technology, San Jose State University,
San Jose, CA 95192, USA

[2] Department of Agricultural and Biosystems Engineering, Iowa State University,
Ames, IA 50011, USA

[3] Machine Learning & Safety Analytics Lab, Department of Technology, San Jose State University,
San Jose, CA 95192, USA

ABSTRACT

Creating safe work environment is significant in saving workers' lives, improving corporates' social responsibility and sustainable development. Pattern identification in occupational accidents is vital in elaborating efficient safety countermeasures aiming at improving prevention and mitigating outcomes of future incidents. The objective of this study is to identify patterns related to the occurrence of occupational accidents in non-farm agricultural work environments based on workers' compensation claims data, using latent class clustering method as an unsupervised machine learning modeling approach. The result

showed injury profiles and incident dynamics have low, average, and high levels of risks based on the main causes and outcomes of the injuries and the affected body part(s).

Keywords: Occupational Accidents, Unsupervised Machine Learning, Latent Class Clustering, Data Analytics, Safety Analytics

INTRODUCTION

Occupational accidents in non-farm agriculture-related work environments have a high rate both in developing and developed countries, despite various safety improvement measures and trainings (Ivascu & Cioca, 2019). Workers in non-farm agribusiness work environments are involved in operating various types of machinery and completing physically-demanding tasks (Cremasco et al., 2019). Grain elevator and feed mill workers are prone to risks such as developing occupational airway and respiratory disease, and incidents from hazardous activities like operating machinery, or fall from different levels.

Analyzing most frequent causes of accidents in agribusiness industries showed that, out of more than 6000 records of incidents in commercial grain elevators in Midwest of the United States, 31% were caused by slip-trip-fall (STF) occurrences. In biofuel agribusiness industry, strain and sprain, laceration, burn and contusion are among the most frequent occupational injuries (Ramaswamy & Mosher, 2018). STF incidents are significant occupational health and safety hazards in various industries including healthcare, manufacturing, retail, and transportation and warehousing (Saadat et al., 2016). In the United States (Bureau of Labor Statistics US Department of Labor, 2018), fatal workplace injuries from slips, trips, and falls have continued a general upward trend, with an increase of 6%, and an overall increment of 25% in the last 10 years (Pomares et al., 2020) and STF incidents and biomechanical ergonomic hazards are the most commonly reported external causes of occupational injuries (Davoudi Kakhki et al., 2021; Meyers et al., 2018). The economic cost from nonfatal work-related fall injuries in the United States was nearly 16 billion USD per year, and over 25% of fall injuries resulted in 31 or more lost workdays (Waehrer et al., 2007; Yoon & Lockhart, 2006). In the United States, STF is one of the most commonly reported external causes of occupational incidents and the percentage of workers' compensation claims ranges from 7% to 44% across industry groups for falling on the same level (Meyers et al., 2018).

Focusing on agribusiness industries, lack of comprehensive database of occupational accidents is a big obstacle in providing accurate evaluation of incident prevalence and patterns (Tolefree et al., 2017). Many studies in occupational safety analytics focus on analyzing structured data using supervised machine learning models and neural networks (Davoudi Kakhki et al., 2019; Ganguli et al., 2021; Kakhki et al., 2019b; Marucci-Wellman et al., 2017; Pishgar et al., 2021; Yedla et al., n.d., 2020). Yet, few focuses on identifying patterns in workplace accidents using unsupervised machine learning modeling techniques,

such as clustering methods, using textual data for analyzing occupational accidents.

Accordingly, the goal of the study was to identify the safety risk profiles of injuries in a population of more than 18,000 agribusiness workers. This study contributes to the limited literature on the use of categorical data to extract meaningful patterns of incidents in safety science, with a focus on agribusiness operations. Furthermore, the analytical approach and results will contribute to informed decision-making with applications in preventing the occurrence or reducing the frequency of occupational incidents among agribusiness workers in non-farm agricultural workplaces.

MATERIAL AND METHODS

We used a record of 18,257 occupational incidents, from 2008 to 2016. Considering the necessity of addressing STF incidents in risk level of occupational injuries, the main causes of injury in this study are divided into two categories: STF incidents and non-STF incidents. The Non-STF category includes causes such as heat or cold exposure, or lifting or handling. Therefore, the distribution is 28.19% for the STF and 71.80% for the Non-STF categories.

The next parameter used is main type of injury, which refers to the injury as being either medical or disability. Another parameter is the part of body that was reported as affected/injured and has seven levels such as upper extremities, neck, etc. The last variable in the study is the agribusiness industry in which the worker was employed, and it includes nine sectors such as commercial grain elevators.

STATISTICAL ANALYSIS USING LATENT CLASS CLUSTERING

Latent Class Analysis (LCA) is an unsupervised machine learning approach that enables identifying qualitatively different classes of input variables in a dataset based on the probabilities of their membership in each class (Weller et al., 2020) , and is limited to modeling categorical (text) variables. Using a probabilistic model, data is divided and classified into a latent class to which their combination has the highest probability or likelihood of belonging. After LCA models are built with a selected range, usually models with 3 to 15 classes, with *k* being the number of estimated parameters in a model, and *n* the number of observations (data points) used in the model, the loglikelihood of each model is calculated, for which Bayesian Information Criterion (BIC) value is gained. BIC values are reliable indicators of model fit, and are used to identify the best number of latent classes that classify the data into meaningful and distinguishable patterns. Lower values of BIC are more desirable and indicate better fits (Morales et al., 2021).

$$BIC = -2LogLikelihood + k\ ln(n)$$

Another consideration in selecting the number of clusters is the interpretability of the

characteristics of each latent class based on literature, theoretical background and applicability of the results in the discipline. Therefore, both statistical criteria and content validity are important regarding selection of the optimal number of latent classes in the final model (Morales et al., 2021). To determine the contributing factors in classifying the data into statistically meaningful groups, the Pearson chi-square statistic (χ^2) is calculated for the contingency table of the expected counts of levels by latent classes. The likelihood ratio test *p*-value for the contingency table of expected counts at $\alpha = 0.01$ significance level is gained and shown as p_{LR}. The $-\log10\ (p_{LR})$ is calculated as the Likelihood Ratio (LR) Logworth statistic. A LR Logworth value above 2 corresponds to being statically significant in differentiating the latent classes at the $\alpha = 0.01$ significant level (because $-\log10\ (0.01) = 2$). Finally, the effect size, which is the contribution of each input variable in distinguishing clusters, is calculated using the chi-squared statistic and the sample size.

RESULTS

The LCC analysis is performed to identify statistically distinctive and meaningful injury profiles of occupational incidents among agribusiness workers based on type of injury, cause of injury, injured body part(s), and agribusiness industries. A total of 13 latent class models are developed with number of clusters set from 3 to 15. When the number of clusters gains four, the lowest BIC is achieved. Therefore, the occupational incidents among workers in agribusiness industries in this study were clustered into four sub-classes. Regarding the probabilities of LCCs, the latent class probabilities of LCC 1, LCC 2, LCC 3, and LCC 4 are 45.12%, 28.49%, 18.60%, and 7.68%, respectively.

As shown in **Error! Reference source not found.**, the information of injury used in the study are statically significant classifiers of latent classes for the selected four-class model, with type of the injury as the most influential factor in segmenting occupational incidents. The incident cause as being STF or Non-STF, the injured body part(s), and the agribusiness industry are the next contributing factors to identifying the injury patterns. Further analysis of conditional probabilities provides a brief description of each LCC and helps to explain the estimated results of the injury profiles among workers.

Table 1: Contributing factors in differentiating LCCs

Variable	Effect size	LR logworth
Main cause of injury	0.7762	2477.7*
Injured body part(s)	0.4818	898.79*
Type of Injury	0.8058	1807.8*
Agribusiness sector	0.3043	333.31*

*Statistically significant classifier of latent clusters at α= 0.05 significance level

In LCC 1, the dominant cause of injury is Non-STF (93.28%), and the major type of injury is medical only (93.48%). As for the most common injured body part(s), such injuries occurred in the upper extremities (53.05%), head (16.44%), trunk (14.67%), and lower extremities (12.71%).

About 50% of the incidents belong to grain elevators and refined fuel industries. LCC 2 is characterized by Non-STF incident cause (86.92) with medical injury outcomes (76.66%), and permanent total/partial disabilities (23.34%). Injuries in trunk and lower extremities account for more than 60% of the injured body part (s) in this category, with work places of grain elevators, refined fuels, and food distributors as the most common ones.

LCC 3 consist of all STF-related injuries (99.18%) that are mainly medical (75.66%), with the highest probability of injuries in lower extremities (41.73%), followed by injuries in trunk and upper extremities (21.21% and 17.73%). Grain elevators and refined fuels are the major agribusiness industries in this class (36.37% and 19.57%).

Incidents in LCC 4 are divided into 61.31% for Non-STF and 38.69% for STF causes, with the main injury type of permanent partial disabilities in 95.16% of the cases. Almost 80% of the injured body part (s) are in upper and lower extremities. The highest probabilities regarding agribusiness sector in LCC 4 belong to grain elevators, feed mills for livestock and refined fuels (48.62%, 13.02%, 12.55%).

The results of the Tukey test show that the mean age of injured workers is statistically different among the four clusters, with LCC 4 having the highest mean age and LCC 1 having the youngest population with average age of 39 years old. Considering the experience years, workers in cluster four have the highest years of experience of average 7 years, followed by cluster 3 with 6 years of experience. The average experience in clusters 1 and 2 are around 5 years and are not statistically different. The details of the analyses are given in **Error! Reference source not found.**.

Table 2: Tukey test for workers' average age and experience (in years)

Cluster	**Mean age**		**Mean experience**	
LCC 4	A	47.08	A	6.94
LCC 3	B	44.946	B	5.95
LCC 2	C	40.89	C	4.91

LCC 1	D	39.86	C	4.84

* Levels not connected by same letter are significantly different at α= 0.05 significance level

The importance of gaining information about the age and experience in the severity of an injury is confirmed by current literature that show severity levels of occupational incidents, and consequently the costs of medical treatment and indemnity, increases with age since it has effects on the physical activity and attention of the workers, specifically in manufacturing operations occupations (Alizadeh et al., 2015; Davoudi Kakhki et al., 2018; Yi, 2018). Regarding the experience, previous literature suggest that reasons behind higher rate of incidents in younger employees are lack of experience and overconfidence (Davoudi Kakhki et al., 2018). Therefore, the average age and experience years of the injured workers in the data are calculated.

Considering the costs of incidents per cluster, LCCs vary based on the average medical, indemnity, and total costs for occupational incidents, as in **Error! Reference source not found.**. Considering the application of the results in safety practices, a big variation in the injury profiles and injury costs results from the main causes of incidents. Our analysis showed that the average total cost of a STF incident is approximately $20,000 while the average total costs is almost half of that value ($10,608) for a non-STF incident, regardless of whether the incident results in either a medical or a disability outcome.

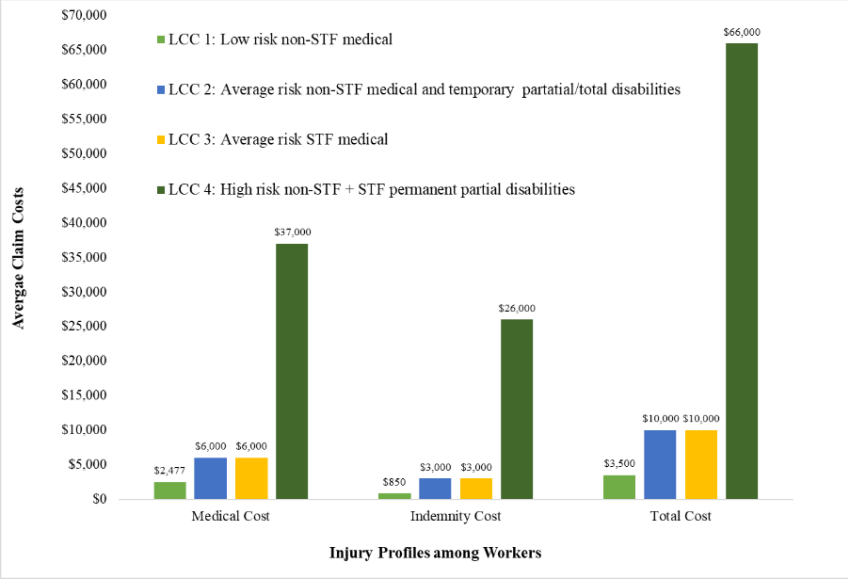

Figure 1: Profiles of occupational accidents among agribusiness workers

The conditional probabilities gained from the analysis of each latent class cluster provide financial and risk managers and safety professionals with information and data to design and implicate preventive measures and strategies both occupation-wise, and industry-wise to achieve the goal of fewer and less severe injuries. If we consider the cost of an incident as a criterion for judging its severity, a medical injury has an average cost of $1440, compared to $22,684 and $67,367 for temporary total/partial disabilities and permanent partial disabilities, respectively. The need for such strategies becomes clearer considering that even though latent class cluster 4 has the least number of occupational incidents, it accounts for the largest proportion of injury costs for medical, indemnity, and total incurred amounts. By reducing the severity of incidents, the costs will consequently decrease, and lives will be saved.

Providing proper ergonomic investment and safety practices in these high-hazards work environments can result in reducing the severity of incident outcomes from any disability type, either temporary partial or total or permanent partial, to only medical outcomes, in which the workers can return to work after the incident without need for taking days away from work. Hence, identifying strategies for reducing the injury severity outcomes can help in estimating health care cost planning and management both for the agribusiness industries employers and insurance companies who provide insurance premium for a client based on their history of prior incident frequency and cost (Davoudi Kakhki et al., 2018).

CONCLUSION

Through the analysis of a large dataset of workers' compensation claims that recorded the history of occupational incidents among agribusiness workers in the Midwest of the United Sates over an eight-year period, we identified four statistically and meaningfully categories of occupational incidents. Based on latent class cluster analysis, we showed that the injury profiles and incident dynamics have low, average, and high levels of risks based on the main causes and outcomes of the injuries and the affected body part(s).

The results suggested that the permanent partial disabilities, resulted from combination of STF and Non-STF incident causes, impose the highest costs for indemnity and medical rehabilitation of the injured workers. Furthermore, the average age and experience years of the workers are the highest among this high-risk profile.

The latent class cluster approach and statistical analysis used in the study can be used by relevant industries for safety risk identification, risk analysis, and informed decision-making process on applying safety measurement plans to avoid potential future incidents. In addition, cluster by cluster analysis can offer insights into the incident dynamics that leads to low, average, or high risks for workers' health and could be useful in risk management evaluation

to prevent more serious health and financial consequences.

REFERENCES

Alizadeh, S. S., Mortazavi, S. B., & Sepehri, M. M. (2015). Analysis of occupational accident fatalities and injuries among male group in Iran between 2008 and 2012. *Iranian Red Crescent Medical Journal*. https://doi.org/10.5812/ircmj.18976

Bureau of Labor Statistics US Department of Labor. (2018). National Census of Fatal Occupational Injuries in 2017. *Bureau of Labor Statistics US Department of Labor*.

Chae, M., & Chung, S. J. (2021). Clustering of south korean adolescents' health-related behaviors by gender: Using a latent class analysis. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph18063129

Cremasco, M. M., Giustetto, A., Caffaro, F., Colantoni, A., Cavallo, E., & Grigolato, S. (2019). Risk assessment for musculoskeletal disorders in forestry: A comparison between RULA and REBA in the manual feeding of a wood-chipper. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph16050793

Davoudi Kakhki, F., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. *Safety Science*, *117*(July 2018), 257–262. https://doi.org/10.1016/j.ssci.2019.04.026

Davoudi Kakhki, F., Freeman, S. A., & Mosher, G. A. (2021). Machine Learning for Occupational Slip-Trip-Fall Incidents Classification Within Commercial Grain Elevators (Vol. 1). *Springer International Publishing*. https://doi.org/10.1007/978-3-030-80288-2_18

Davoudi Kakhki, F., Freeman, S., & Mosher, G. (2018). Analyzing Large Workers' Compensation Claims Using Generalized Linear Models and Monte Carlo Simulation. *Safety*, *4*(4), 57. https://doi.org/10.3390/safety4040057

Ganguli, R., Miller, P., & Pothina, R. (2021). Effectiveness of Natural Language Processing Based Machine Learning in Analyzing Incident Narratives at a Mine. *Minerals* https://doi.org/10.3390/min11070776

Ivascu, L., & Cioca, L. I. (2019). Occupational accidents assessment by field of activity and investigation model for prevention and control. *Safety*. https://doi.org/10.3390/safety5010012

Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2019a). Use of logistic regression to identify factors influencing the post-incident state of occupational injuries in agribusiness operations. *Applied Sciences (Switzerland)*. https://doi.org/10.3390/app9173449

Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2019b). Use of neural networks to identify safety prevention priorities in agro-manufacturing operations within commercial grain elevators. *Applied Sciences (Switzerland)*. https://doi.org/10.3390/app9214690

Marucci-Wellman, H. R., Corns, H. L., & Lehto, M. R. (2017). Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review. *Accident Analysis and Prevention*, *98*, 359–371. https://doi.org/10.1016/j.aap.2016.10.014

Meyers, A. R., Al-Tarawneh, I. S., Wurzelbacher, S. J., Bushnell, P. T., Lampl, M. P., Bell, J. L., Bertke, S. J., Robins, D. C., Tseng, C. Y., Wei, C., Raudabaugh, J. A., & Schnorr, T. M. (2018). Applying machine learning to workers' compensation data to identify industry-specific ergonomic and safety prevention priorities: Ohio, 2001 to 2011. *Journal of Occupational and Environmental Medicine*. https://doi.org/10.1097/JOM.0000000000001162

Morales, A., Melero, S., Tomczyk, S., Espada, J. P., & Orgilés, M. (2021). Subtyping of Strengths and Difficulties in a Spanish Children Sample: A Latent Class Analysis. *Journal of Affective Disorders*. https://doi.org/10.1016/j.jad.2020.11.047

Pishgar, M., Issa, S. F., Sietsema, M., Pratap, P., & Darabi, H. (2021). Redeca: A novel framework to review artificial intelligence and its applications in occupational safety and health. In *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph18136705

Pomares, J. C., Carrión, E. Á., González, A., & Saez, P. I. (2020). Optimization on personal fall arrest systems. Experimental dynamic studies on lanyard prototypes. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph17031107

Ramaswamy, S. K., & Mosher, G. A. (2018). Using workers' compensation claims data to characterize occupational injuries in the biofuels industry. *Safety Science*. https://doi.org/10.1016/j.ssci.2017.12.014

Saadat, S., Hafezi-Nejad, N., Ekhtiari, Y. S., Rahimi-Movaghar, A., Motevalian, A., Amin-Esmaeili, M., Sharifi, V., Hajebi, A., Radgoodarzi, R., Hefazi, M., Eslami, V., Karimi, H., Mohammad, K., & Rahimi-Movaghar, V. (2016). Incidence of fall-related injuries in Iran: A population-based nationwide study. *Injury*. https://doi.org/10.1016/j.injury.2016.05.001

Swaen, G. M. H., Van Amelsvoort, L. G. P. M., Bültmann, U., & Kant, I. J. (2003). Fatigue as a risk factor for being injured in an occupational accident: Results from the Maastricht Cohort Study. *Occupational and Environmental Medicine*. https://doi.org/10.1136/oem.60.suppl_1.i88

Tolefree, S., Truong, A., Ward, J., Dong, F., Ablah, E., & Haan, J. (2017). Outcomes Following Traumatic Grain Elevator Injuries. *Journal of Agromedicine*. https://doi.org/10.1080/1059924X.2017.1318727

Unsar, S., & Sut, N. (2009). General assessment of the occupational accidents that occurred in Turkey between the years 2000 and 2005. *Safety Science*. https://doi.org/10.1016/j.ssci.2008.08.001

Waehrer, G. M., Dong, X. S., Miller, T., Haile, E., & Men, Y. (2007). Costs of occupational injuries in construction in the United States. *Accident Analysis and Prevention*. https://doi.org/10.1016/j.aap.2007.03.012

Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent Class Analysis: A Guide to Best Practice. *Journal of Black Psychology*. https://doi.org/10.1177/0095798420930932

Yedla, A., Kakhki, F. D., & Jannesari, A. (2020). Predictive modeling for occupational safety outcomes and days away from work analysis in mining operations. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph17197054

Yi, K. H. (2018). The High-risk Groups According to the Trends and Characteristics of Fatal Occupational Injuries in Korean Workers Aged 50 Years and Above. *Safety and Health at Work*. https://doi.org/10.1016/j.shaw.2018.01.005

Yoon, H. Y., & Lockhart, T. E. (2006). Nonfatal occupational injuries associated with slips and falls in the United States. *International Journal of Industrial Ergonomics*. https://doi.org/10.1016/j.ergon.2005.08.005