Joint CSCE Construction Specialty & CRC Conference 2025
Conférence conjointe spécialisée en construction de la SCGC et CRC-2025

Montreal, Quebec
July 28-31, 2025 / *28-31 juillet 2025*

Analyzing Significant Keywords Relationships in Construction Safety Based on OSHA's Accident Reports

Jeon, B.J.[1] and Lee, H.W.[1]*

[1] Department of Construction Management, University of Washington, Seattle, WA 98195, USA

ABSTRACT: The construction industry is among the most hazardous sectors with injuries often resulting in severe consequences, including life-long disabilities and fatalities. Over the years, various strategies have been developed to improve worker safety and accident prevention. While recent advancements in Natural Language Processing (NLP) have enabled researchers to derive meaningful insights from construction-related reports, little to no studies have specifically aimed to apply NLP to analyze accident reports. In response, this study aims to apply Term Frequency-Inverse Document Frequency (TF-IDF) and frequency analyses to examine accident data from the U.S. Occupational Safety and Health Administration (OSHA). The specific study objectives are (1) to identify significant accident-related keywords by North American Industry Classification System (NAICS) codes and (2) to explore the contextual relationships between NAICS codes and keywords. The analysis revealed that fall and equipment-related accidents were predominant across the construction industry. Additionally, certain NAICS codes showed unique categories such as ELECTRICITY, BURN, and CAVE IN. These findings provide valuable insights into the relationship between keywords and construction types, equipping safety managers and project stakeholders with a deeper understanding of factors influencing construction safety. This enhanced understanding will serve as a foundation for reviewing safety policies tailored to specific construction types, ultimately enabling the development of targeted prevention strategies.

1. INTRODUCTION

The construction industry is one of the most hazardous sectors, with a disproportionately high rate of fatal accidents and injuries compared to other industries. This heightened risk is attributed to construction activity's complex and dynamic nature, which often involves heavy machinery, working at heights, and exposure to various environmental hazards (Almaskati et al. 2024; Kang 2025; Nnaji et al. 2020; Osburn et al. 2022). However, the construction sector is indispensable, forming the backbone of modern infrastructure and significantly contributing to economic development and societal well-being. Its potential for growth and innovation is immense, making it a critical area for research and improvement, particularly in safety and efficiency (Cao et al. 2025; Chen and Cao 2021).

Furthermore, the economic implications of construction accidents are profound. Studies have shown a negative correlation between GDP growth and workplace fatalities, suggesting that economic development alone cannot address safety challenges without targeted interventions (Cao et al. 2025). The construction industry's role in driving economic growth further amplifies the need for sustainable safety practices that protect workers while maintaining productivity.

Thus, recent studies have continued highlighting the persistent challenges in construction safety, emphasizing the critical need for advanced risk mitigation methodologies. For instance, research utilizing

data envelopment analysis (DEA) models has demonstrated significant regional disparities in safety performance across China, with some provinces showing a 70% potential reduction in construction fatalities annually (Kang 2025). Similarly, machine learning approaches, such as graph convolutional networks (GCN), have been employed to accurately predict accident outcomes, offering insights into human and workplace factors contributing to safety incidents (Mostofi and Toğan 2024). These advancements underscore the importance of integrating technological and analytical tools to enhance safety protocols and reduce fatalities.

While the construction industry is inherently risky, its critical role in societal development and its innovation potential necessitates advanced research into safety and efficiency. Previous studies have attempted to address these challenges by applying methodologies such as NLP to achieve this. However, despite these efforts, there is still a lack of research specifically focused on analyzing accident reports. This study aims to identify key trends and provide actionable insights to enhance safety protocols by leveraging analytical tools, such as keyword analysis of construction accident reports based on North American Industry Classification System (NAICS) codes. The industry can mitigate risks and achieve sustainable growth by fostering a safety culture and implementing data-driven strategies. This paper contributes to the existing body of knowledge by offering a detailed analysis of construction accident patterns to improve safety and efficiency in the construction sector.

2. LITERATURE REVIEW

2.1 Construction Safety

The construction industry is one of the most hazardous sectors globally, with high accident rates and significant safety challenges. Recent studies have extensively explored the causes, prevention strategies, and technological advancements in construction safety. For instance, Namian et al. (2022) highlights the differences in safety perceptions between construction managers and workers, emphasizing the need for improved communication and training to mitigate risks. Similarly, Karakhan and Gambatese (2018) discusses the role of incentives and rewards in enhancing safety outcomes, suggesting that behavioral and organizational factors play a critical role in accident prevention. Other research, such as that by Jeelani et al. (2017), identifies the persistent issue of unrecognized hazards at construction sites, advocating for more robust hazard identification methodologies. These studies collectively underscore the importance of addressing human and systemic factors to improve safety performance in construction.

Technological innovations have increasingly complemented these behavioral strategies. Alizadehsalehi et al. (2020) demonstrated the integration of Building Information Modeling (BIM) with unmanned aerial vehicles (UAVs) for real-time safety monitoring. Zhong et al. (2020) proposed ontology-based models to extract hazards from construction imagery. More recently, wearable sensing devices and AI-driven systems have been explored for proactive and personalized safety management (Nnaji et al., 2021; Zhou and Zhang, 2023). Chandu et al. (2024) highlighted how combining wearable technology, artificial intelligence, and digital training could transition safety practices from reactive to proactive approaches. In parallel, organizational and cultural dimensions have also been emphasized: for instance, Meng and Chan (2022) demonstrated that individual perception and organizational collectivity jointly shape safety outcomes in practice.
These trends underscore a growing convergence of technological, behavioral, and organizational approaches in modern construction safety management.

2.2 Natural Language Process

Natural Language Processing (NLP) has emerged as a powerful tool in the construction industry, particularly in enhancing safety management. Researchers have explored various applications of NLP to address challenges such as document classification, risk assessment, and compliance checking. For instance, Zhang et al. (2015) demonstrated the use of NLP in automating the classification of construction project documents, significantly improving efficiency and reducing manual errors. Li et al. (2024) proposed an NLP-based framework integrated with knowledge graphs to automate compliance checking for BIM

models, enabling efficient identification of errors and reducing legal risks. These advancements highlight the potential of NLP in transforming unstructured text data into actionable insights, thus improving decision-making processes in construction safety management.

Recent studies have also focused on leveraging NLP for risk prediction and safety monitoring. For example, Tixier et al. (2016) utilized NLP to analyze safety reports and identify patterns that could predict potential hazards in construction sites. Similarly, Zhang and El-Gohary (2016) developed an NLP-based framework to extract and analyze safety-related information from construction documents, laying the groundwork for proactive risk management. These studies underscore the versatility of NLP in addressing diverse safety-related challenges in the construction industry. More recent developments in large language models (LLMs), such as BERT and GPT, offer promising capabilities for deeper semantic understanding of incident narratives and risk classification, though their application in construction safety remains limited (Smentana et al. 2024; Saka et al. 2024).

While NLP has been increasingly applied to construction safety management, there is limited research on leveraging NLP to perform keyword analysis based on NAICS codes. Existing studies have primarily focused on document classification, compliance checking, and risk prediction, yet little to no studies have systematically analyzed accident data to extract meaningful insights specific to different industry sectors defined by NAICS codes. This limitation hinders identifying sector-specific safety trends and risks critical for developing targeted safety interventions. In response, focusing on keyword analysis aligned with NAICS codes, this study aims to address this gap and provide a data-driven approach to understanding safety challenges across diverse construction sectors.

3. RESEARCH METHODOLOGY

3.1 Data Collection

This study involves collectting accident case data from OSHA, identifying key keywords by type of construction work through keyword analysis, comparing them, and identifying accident-inducing keywords and characteristics of construction activities according to NAICS codes. To achieve this, data was collected from OSHA's accident report pages under NAICS code 23 (construction sector) using web crawling, an automated method that efficiently gathers large-scale data by navigating web pages. The Python library BeautifulSoup was utilized for this process, resulting in a total of 27,598 cases.

During data preprocessing, the dataset was filtered and organized according to the number of employees involved in each incident, and irrelevant data entries were removed to ensure accurate analysis. However, the keywords themselves were not modified, as they were directly extracted from OSHA's predefined categories. This approach preserved the integrity of OSHA's keyword definitions while optimizing the dataset for analysis. The collected data is shown in Table 1.

Table 1: OSHA Dataset: Data Elements Overview

Summary	Keyword	Inspection No.	Date	NAICS	Employ No.	Age	Sex	Degree of Injury

Each report contains detailed information, including the circumstances of the accident, employee and employer information, and the nature and location of injuries. However, for this study, only filtered dataset in Table 2 was used after preprocessing

Table 2: Filtered Dataset: Selected Data Elements for Analysis

Keyword	Inspection No.	Date	NAICS	Employ No.

To highlight the significance of accidents involving multiple employees, cases in which multiple employees were injured in a single accident were recorded separately. For instance, if three employees were injured

in one accident, three cases with the same keyword, date, and NAICS code but different employee numbers (Employ No.) were documented. This resulted in a total of 30,234 cases.

Additionally, when organizing the data by year, cases before 2003 and after 2024, which had fewer than 300 cases, were excluded. The analysis was limited to data from 2004 to 2023, resulting in a final dataset of 27,683 cases. The following Figure 1 illustrates the trend in the number of accidents by year

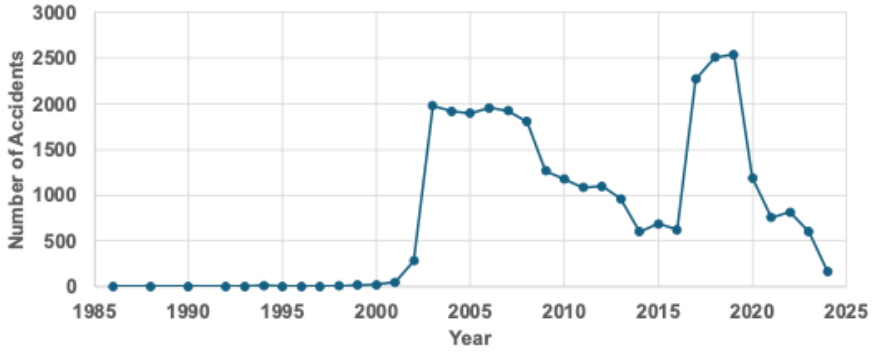

Figure 1: Number of Accidents Over Time

The graph reveals notable inflection points in 2009, 2016, and 2020. The declines observed in 2009 and 2020 coincide with economic recessions triggered by the Subprime Mortgage Crisis and the COVID-19 pandemic, both of which led to substantial reductions in construction activity (Venugopal, M. 2020; Pamidimukkala, A. and Kermanshachi, S. 2021). The subsequent rise in 2017 appears to be driven by a combination of factors, including a construction boom that brought an influx of inexperienced workers and a reduction in OSHA's compliance inspection workforce.

From a methodological standpoint, changes to OSHA's electronic reporting requirements between 2016 and 2019 may have led to a more comprehensive inclusion of non-fatal incidents. This reporting modification could affect longitudinal comparability and should be carefully considered when interpreting trends in construction accident data across these years (Bluegreen Alliance 2024).

3.2 Data Analysis

To evaluate the importance of keywords by NAICS code, both frequency analysis and TF-IDF (Term Frequency-Inverse Document Frequency) methods were employed. The Frequency Analysis counts the absolute frequency of specific words within each NAICS code. The NLTK(Natural Language Toolkit) library was used to tokenize the text and count the occurrences of each word. However, frequency analysis alone has limitations, as it may overemphasize common words that appear frequently across documents.

TF-IDF is a method used in information retrieval and text mining to evaluate the importance of a word within a document set. It combines two factors: Term Frequency (TF), which measures how often a word appears in a document, and Inverse Document Frequency (IDF), which measures how rare a word is across all documents. The Term Frequency (TF) is calculated as follows:

$$[1]\ \mathrm{TF}(t, d) = \frac{\text{Number of times term t apears in document d}}{\text{Total number of terms in document d}}$$

The Inverse Document Frequency (IDF) is calculated as:

$$[2]\ \mathrm{IDF}(t, d) = \log\left(\frac{N}{df(t)}\right)$$

where N is the total number of documents in the dataset, and df(t) is the number of documents containing the term t.

The TF-IDF value is calculated by multiplying TF and IDF, as shown in the following formula:

[3] $\mathrm{TF-IDF(t, d, D)} = \mathrm{TF(t, d)} \times \mathrm{IDF(t, D)}$

The Python library scikit-learn was used to perform TF-IDF analysis on the keywords that were grouped using the NAICS code.

3.3 Result

The results of the frequency analysis are summarized in Table 3. The most frequently occurring keywords across all NAICS codes were related to falls (e.g., FALL, FRACTURE, HEAD) and equipment-related incidents (e.g., STRUCK BY). The NAICS codes with the highest number of keywords were Roofing Contractors (238160), Electrical Contractors and Other Wiring Installation Contractors (238210), and Commercial and Institutional Building Construction (236220). These codes also showed a high frequency of fall-related keywords. However, in Electrical Contractors and Other Wiring Installation Contractors (238210), keywords related to electrical incidents (e.g., ELECTRICAL WORK, ELECTRICIAN, BURN) were also prominent.

Keywords in Tables 3 and 4 are color-coded based on their semantic categories: yellow denotes incident types or causes (e.g., fall, collapse), green represents work environments or associated objects (e.g., construction site, ladder), and cyan highlights worker roles, activities, or affected body parts (e.g., installing, head injury).

Table 3: Frequency Analysis of Keywords by NAICS Code (Subset of Total Data)

RANK	Commercial and Institutional Building Construction (236220)	Water and Sewer Line and Related Structures Construction (237110)	Highway, Street, and Bridge Construction (237310)	Roofing Contractors (238160)	Electrical Contractors and Other Wiring Installation Contractors (238210)	Overall Rank
1	FALL	CONSTRUC-TION	STRUCK BY	FALL	CONSTRUC-TION	FALL
2	CONSTRUC-TION	STRUCK BY	CONSTRUC-TION	ROOF	FALL	CONSTRUC-TION
3	FRACTURE	TRENCH	FRACTURE	FALL PROTECTION	ELECTRICAL WORK	FRACTURE
4	STRUCK_BY	FRACTURE	FALL	ROOFER	FRACTURE	STRUCK BY
5	FALL PROTECTION	EXCAVATION	HIGHWAY	FRACTURE	ELECTRICIAN	FALL PROTECTION
6	HEAD	CRUSHED	MOTOR VEHICLE	CONSTRUC-TION	ELECTRICAL	HEAD
7	ROOF	COLLAPSE	CRUSHED	HEAD	INSTALLING	ROOF
8	FALLING OBJECT	FALL	RUN_OVER	LADDER	BURN	INSTALLING
9	LEG	PIPE	HEAD	INSTALLING	ELECTRO-CUTED	LADDER
10	COLLAPSE	ASPHYXIAT-ED	TRUCK	RESIDENTIAL CONSTRUC-TION	ELECTRIC SHOCK	LEG

In Highway, Street, and Bridge Construction (237310), equipment-related keywords (e.g., STRUCK_BY, CRUSH, RUN OVER) appeared twice as frequently as fall-related keywords. Similarly, in Water and Sewer Line and Related Structures Construction (237110), equipment-related keywords (e.g., STRUCK_BY, TRENCH) were more frequent than fall-related keywords, suggesting a higher prevalence of equipment-related accidents in these sectors.

Table 4: TF-IDF Analysis of Keywords by NAICS Code (Subset of Total Data)

RANK	Commercial and Institutional Building Construction (236220)	Water and Sewer Line and Related Structures Construction (237110)	Highway, Street, and Bridge Construction (237310)	Roofing Contractors (238160)	Electrical Contractors and Other Wiring Installation Contractors (238210)	Overall Rank
1	FALL	CONSTRUC-TION	STRUCK BY	FALL	CONSTRUC-TION	FALL
2	CONSTRUC-TION	STRUCK BY	CONSTRUC-TION	ROOF	FALL	CONSTRUC-TION
3	FRACTURE	TRENCH	HIGHWAY	ROOFER	ELECTRICIAN	FRACTURE
4	STRUCK_BY	EXCAVA-TION	FRACTURE	FALL PROTECTION	ELECTRICAL WORK	STRUCK BY
5	FALL PROTECTION	FRACTURE	ROAD PAVING	FRACTURE	BURN	FALL PROTEC-TION
6	HEAD	CRUSHED	FALL	CONSTRUC-TION	FRACTURE	ROOF
7	ROOF	PIPE	MOTOR VEHICLE	HEAD	ELECTRICAL	HEAD
8	FALLING OBJECT	COLLAPSE	CRUSHED	LADDER	INSTALLING	INSTALLING
9	LEG	SEWER	RUN_OVER	INSTALLING	ELECTRO-CUTED	LADDER
10	INSTALLING	CAVE IN	BRIDGE	RESIDENTIAL CONSTRUC-TION	ELECTRIC SHOCK	LEG

The TF-IDF results were generally consistent with the frequency analysis, but some differences emerged. For example, in Electrical Contractors and Other Wiring Installation Contractors (238210), the keyword BURN, which ranked 8th in frequency analysis, rose to 5th in TF-IDF analysis. This suggests that burn-related incidents are particularly significant in electrical work. Similarly, in Highway, Street, and Bridge Construction (237310) and Water and Sewer Line and Related Structures Construction (237110), keywords like HIGHWAY and CAVE IN rose in rank, reflecting the unique risks associated with these types of construction work.

4. CONCLUSIONS

This study compared the characteristics of keywords in OSHA accident reports using frequency analysis and TF-IDF analysis. The results showed that fall-related keywords (e.g., FALL, FRACTURE, HEAD) and equipment-related keywords (e.g., STRUCK BY, CRUSH) were the most frequent across all NAICS codes. However, specific NAICS codes exhibited unique keyword patterns. For example, Electrical Contractors and Other Wiring Installation Contractors (238210) showed a high frequency of ELECTRICAL and BURN, while Highway, Street, and Bridge Construction (237310) and water and sewer construction (237110) had a higher prevalence of equipment-related keywords like STRUCK_BY and TRENCH.

The comparison between frequency analysis and TF-IDF analysis revealed that while both methods produced similar keyword rankings, TF-IDF was determined to be more effective in highlighting keywords that reflect the unique characteristics of specific construction types. For instance, BURN in electrical work and HIGHWAY and CAVE_IN in highway and water/sewer construction were more prominent in the TF-IDF results, indicating their importance in these sectors.

The findings of this study establish a foundational dataset for analyzing construction-type-specific safety risks and provide actionable insights for on-site project managers. While fall and equipment-related accidents are ubiquitous across construction sectors, their prevalence and characteristics exhibit NAICS code-specific variations. Notably, the TF-IDF analysis identified that BURN and ELECTRICAL WORK

showed statistically significant prominence in electrical construction, whereas STRUCK BY and TRENCH were more prevalent in highway and utility-related projects. These insights enable managers to pinpoint critical incident types within their operational domains and develop risk-specific mitigation strategies. Furthermore, keyword-based analysis supports prioritizing safety training, inspection protocols, and resource allocation by aligning preventive measures with data-driven risk profiles. Therefore, this study advances both academic knowledge and evidence-based practices in construction safety management

A limitation of this study is the restricted scope of data, as it focused on a specific industry sector and used predefined keywords from OSHA. Future research should expand the analysis to include a wider range of industries and larger datasets. Additionally, the effectiveness of safety policies based on keyword analysis should be evaluated in subsequent studies

ACKNOWLEDGMENTS

This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2024-00410520).

REFERENCES

Alizadehsalehi, S., I. Yitmen, T. Celik, and D. Arditi. 2020. The effectiveness of an integrated BIM/UAV model in managing safety on construction sites. *International Journal of Occupational Safety and Ergonomics*, 26 (4): 829–844.

Almaskati, D., S. Kermanshachi, A. Pamidimukkala, K. Loganathan, and Z. Yin. 2024. A Review on Construction Safety: Hazards, Mitigation Strategies, and Impacted Sectors. *Buildings*, 14 (2): 526.

BlueGreen Alliance. 2024. Then and Now: Worker Safety under Trump and Biden. BlueGreen Alliance, Minneapolis, MN, USA.

Cao, Z., T. Zhou, S. Miao, L. Wang, and Z. Wang. 2025. Exploring the economic occupational health, safety, and fatal accidents in high-risk industries. *BMC Public Health*, 25 (1): 433.

Chandu, K.P., Raja, K.H., and Kumar, N.N. 2024. From Reactive to Proactive: The Role of Wearable Technology, AI, and Digital Training in Construction Safety Management. *Library Progress International*, 44(3): 22858-22864.

Chen, T., and Z. Cao. 2021. Construction safety: an analysis of the cross-influence of economic, construction, and accident death factors. *Environ Sci Pollut Res*, 28 (46): 65243–65254.

Jeelani, I., A. Albert, and J. A. Gambatese. 2017. Why Do Construction Hazards Remain Unrecognized at the Work Interface? *J. Constr. Eng. Manage.*, 143 (5): 04016128.

Kang, L. 2025. Examining safety conditions in the construction sector across Chinese provinces: an input-output analysis. *Environment, Development and Sustainability*, 1-23.

Karakhan, A., & Gambatese, J. 2018. Hazards and risk in construction and the impact of incentives and rewards on safety outcomes. *Practice Periodical on Structural Design and Construction*, *23*(2), 04018005.

Li, S., Wang, J., & Xu, Z. 2024. Automated compliance checking for BIM models based on Chinese-NLP and knowledge graph: an integrative conceptual framework. *Engineering, Construction and Architectural Management*.

Mostofi, F., and V. Toğan. 2024. Predicting Construction Accident Outcomes Using Graph Convolutional and Dual-Edge Safety Networks. *Arab J Sci Eng*, 49 (10): 13315–13332.

Meng, X. and Chan, A.H.S. 2022. Improving the Safety Performance of Construction Workers through Individual Perception and Organizational Collectivity: A Contrastive Research between Mainland China and Hong Kong. *International Journal of Environmental Research and Public Health*, 19(21): 14599

Namian, M., M. Tafazzoli, A. J. Al-Bayati, and S. Kermanshachi. 2022. Are Construction Managers from Mars and Workers from Venus? Exploring Differences in Construction Safety Perception of Two Key Field Stakeholders. *International Journal of Environmental Research and Public Health*, 19 (10): 6172.

Nnaji, C., Karakhan, A. A., Gambatese, J., & Lee, H. W. 2020. Case study to evaluate work-zone safety technologies in highway construction. *Practice Periodical on Structural Design and Construction*, *25*(3), 05020004.

Nnaji, C., Awolusi, I., Park, J., and Albert, A. 2021. Wearable Sensing Devices: Towards the Development of a Personalized System for Construction Safety and Health Risk Mitigation. Sensors, 21(3): 682.

Osburn, L., Lee, H. W., & Gambatese, J. A. 2022. Formal prevention through design process and implementation for mechanical, electrical, and plumbing worker safety. *Journal of Management in Engineering*, *38*(5), 05022011

Pamidimukkala, A. and Kermanshachi, S. 2021. Impact of Covid-19 on Field and Office Workforce in Construction Industry. *Project Leadership and Society*, 2: 100018.

Saka, N., Taiwo, R., Salami, B.A., Ajayi, S., Akande, K., and Kazemi, H. 2024. GPT Models in Construction Industry: Opportunities, Limitations, and a Use Case Validation. *Developments in the Built Environment*, 17: 100300.

Smetana, M., Salles de Salles, L., Sukharev, I., and Khazanovich, L. 2024. Highway Construction Safety Analysis Using Large Language Models. *Applied Sciences*, 14: 1352.

Tixier, A. J.-P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016. Application of machine learning to construction injury prediction. *Automation in Construction*, 69: 102–114.

Venugopal, M. 2020. Construction Fatalities in the United States Between 2009–2018. *International Journal for Research in Applied Science and Engineering Technology*, 8(IV): 1129–1133.

Zhang, J., and N. M. El-Gohary. 2016. Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking. *J. Comput. Civ. Eng.*, 30 (2): 04015014.

Zhang, S., F. Boukamp, and J. Teizer. 2015. Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for job hazard analysis (JHA). *Automation in Construction*, 52: 29–41.

Zhong, B., H. Li, H. Luo, J. Zhou, W. Fang, and X. Xing. 2020. Ontology-Based Semantic Modeling of Knowledge in Construction: Classification and Identification of Hazards Implied in Images. *J. Constr. Eng. Manage.*, 146 (4): 04020013.

Zhou, Y. and Zhang, M. 2023. The Impact of Wearable Devices on the Construction Safety of Building Workers: A Systematic Review. *Sustainability*, 15(14): 11165.