MACHINE LEARNING-DRIVEN ACCIDENT PREDICTION TO ENHANCE INHERENT SAFETY DESIGN THROUGHOUT PROCESS DESIGN LIFECYCLE

HASLINDA BINTI ABDUL SAHAK
MCS241004_MCST 1043
PROPOSAL

UNIVERSITI TEKNOLOGI MALAYSIA

NOVEMBER 2024

SCHOOL OF COMPUTING
Faculty of Engineering

Project Proposal Form MCSD 6215
Sem:..1.......... Session:..2024 / 2025......

SECTION A: Project Information.

Program Name:	**Masters of Data Science**
Subject Name:	**Project 1 (MCSD 6215)**
Student Name:	Haslinda binti Abdul Sahak
Metric Number:	MCS241004
Student Email & Phone:	haslinda45@graduate.utm.my & 011-7035 8229
Project Title:	Machine Learning-Driven Accident Prediction to Enhance Inherent Safety Design throughout Process Design Lifecycle
Supervisor 1:	
Supervisor 2 / Industry Advisor (if any):	

SECTION B: Project Proposal

Introduction:

Inherent safety design (ISD) is a forward-thinking way of preventing accidents and minimizing risks throughout the design life cycle of any process. Industries around the globe have been evolving and yet the most important factor of all is whether or not an operation is safe, more particularly in high-risk industries such as manufacturing, oil and gas, and chemicals. Most safety designs are based on human judgment, historical data, and manual intervention. All this inaccuracy leads to inefficiency. The prediction of potential accidents has entered a new era with machine learning (ML). This makes the whole process of making designs with inherent safety even more feasible. The insights from pattern identification will provide inputs into safety design before accidents occur using predictive algorithms created with large datasets. This research, hence, will look into and actualize the impact of ML in accident prediction to further improve inherent safety design optimization along the process design lifecycle, thus reducing the accident risk and enhancing safety performance.

Problem Background:

There are various phases of the process design lifecycle, such as conception and operation, during this period wherein safety should be a concern. Safety design today has mostly been reactive, with most accidents occurring before preventive measures against the same could be taken. More importantly, regardless of all the existing standards and regulations to govern accident prevention, many accidents still occur because certain risks are considered or underrated. The current safety measures still use a well-proven but static predictive approach. Machine learning asks about the possibility of making inherent safety design better by forecasting accidents through data and insights into hidden risks for safety measures optimization throughout the design lifecycle. As yet, applying ML in safety design is still in its infancy, and several challenges continue to exist in integrating predictive models with real-world process design systems.

Problem Statement:

Traditional safety design methods tend to be reactive, often failing to predict accidents before they occur. This results in unsafe conditions and unnecessary risks. Additionally, the current approach to inherent safety design does not fully utilize machine learning and predictive analytics, which could help identify risk patterns early in the process lifecycle. Furthermore, there is a lack of standardized methodologies for integrating machine learning-driven accident prediction into the process design lifecycle, leading to inconsistent safety measures across industries.

Despite the comprehensive safety information and preventive measures available, the chemical process industry (CPI) witnesses a high number of accidents. A significant factor contributing to this is the ineffectiveness in communicating the lessons of accidents which makes learning from accidents obscure - that is, translating lessons into actions, changing one's conduct in order not to repeat a given mistake, or acquiring knowledge that can be used as a safeguard against reoccurrences in the future. Appropriate, applicable, effective, and efficient methods exist for the design of safety and health in CPI systems but they are rarely utilized in an optimum manner or even at all, are application specific, complex and long duration activities, and require complex interpretation that most people do not have the knowledge or the experience of carrying out. Most of these methods are on the other hand appropriate for the late stage of development and have low use in the early stage when the design itself has little information.

Given these challenges, it is crucial to implement a comprehensive, data-focused strategy that incorporates past accidents to enhance safety and health during the entire design process. This method aims to use accident databases to gather and examine data on different types of accidents, their causes, and ways to prevent them, creating an organized and easy-to-access source of safety knowledge that can be used in every stage of design. Through the integration of machine learning for data analysis, it becomes feasible to forecast potential hazards at an early stage in the design process and improve accident understanding. This study aims to create a data-driven system that systematically combines fundamental safety and health principles into CPI design, with the ultimate goal of reducing accident rates and enhancing safety within the industry.

Aim of the Project:

This research aims to develop a machine learning and prediction framework for data-driven analysis of accident-based information to enhance inherent safety and health throughout the process design life cycle in the chemical process industry (CPI). This framework will utilize historical accident data to identify, classify, and rank accident contributors and safety recommendations, providing actionable insights that support proactive, informed safety and health decision-making from early design phases through the process design life cycle. Ultimately, this research aims to reduce accident recurrence by embedding safety and health knowledge into CPI design practices.

Objectives of the Project:

The primary objective of this research is to enhance accident learning in the chemical process industry (CPI) by safety and health insights into process development and design. Specifically, the research aims:

1. To identify, classify, and rank key accident contributors and process safety recommendations throughout the process design lifecycle.
2. To analyze accident database information for inherent safety design, applying insights to:
   a. Research and Development (R&D)
   b. Preliminary Engineering Phase
   c. Basic Engineering Phase
   d. Detailed Engineering Phase
   e. Procurement, Fabrication, Commissioning and Start-up
   f. Operation/ Plant modification
3. To interconnect accident contributors against process safety hierarchy and determine the current preference of loss prevention in the industry.

Scopes of the Project:

The scope of this research is to analyze a large volume of historical accident data and categorize insights relevant to safety and health integration across process design stages in the chemical process industry (CPI). The scope is defined as follows:

1. Data Collection: The research will retrieve at least 500,000 accident investigation reports from reputable sources, including the US Chemical Safety and Hazard Investigation Board (CSB), the US National Transportation Safety Board (NTSB), the US Environmental Protection Agency (EPA), Japan's Science and Technology Failure Knowledge Database (FKD), and the EU Major Accident Reporting System (e-MARS). These reports cover accidents from 1990 to 2024.

2. Accident Analysis: The study will focus on process safety-related accidents to identify critical contributors in the CPI. These contributors will be classified by design error categories, such as process condition, reactivity/incompatibility, unsuitable equipment/part, protection, construction material, layout, utility setup, sizing, and automation/instrument issues, all linked to equipment failure types.

3. Classification of Accident Contributors and Root Causes: Identified contributors will be categorized into design, technical, human, organizational, and external errors. Root causes will be further classified by process design life cycle: research and development (R&D), preliminary engineering, basic engineering, detailed engineering, procurement, fabrication, commissioning start-up, and operation/ plant modification.

4. Recommendations Analysis: Recommendations within the reports will be classified based on preventive, protective, mitigative, and corrective actions. These safety actions are further categorized by inherent safety, , passive-engineered, and active-engineered measures. Inherent safety and health recommendations will focus on ISD (Inherently Safer Design) principles: minimization, substitution, moderation, and simplification.

5. Statistical Ranking: Statistical analysis will classify and rank the accident contributors and safety recommendations based on frequency of occurrence, allowing the identification of critical design-related issues in CPI accidents.

6. Interconnection analysis of the contributors against process safety actions is conducted on single-cause accidents and overall accidents. For interconnection analysis of the overall accidents, a contributor-action mapping using brainstorming among researchers is used to identify the corresponding action(s) of each contributor as the reports generally list their recommendations.

.

Expected Contribution of the Project:

This research is expected to have a significant impact on safety and health procedures in the chemical process sector through the development of a strong machine-learning framework for accident data analysis. The project will offer a systematic method for incorporating accident information into process development, allowing for the early detection of potential risks and successful safety measures. This framework will assist designers in making informed safety decisions throughout the engineering process by categorizing and ranking accident contributors and recommendations. Additionally, the project seeks to enhance the accessibility and utilization of accident information through improved data sharing and encouraging continual learning from previous events. This framework based on data will be beneficial for academic research and industry practices, helping to make process design safer and possibly lower accident rates in the CPI.

Project Requirements:

Software:	1. **Data Processing & Analysis**: Use *Python* with libraries like *Pandas*, *NumPy*, and *scikit-learn* to handle, analyze, and model accident data 2. **Database Management**: Use *SQL* (like *MySQL*) for structured data or *NoSQL* (like *MongoDB*) for more flexible data storage and *Apache Spark* for large-scale data processing if needed. 3. **Machine Learning Frameworks:** Use *TensorFlow* or *PyTorch* to build predictive models and *MLflow* to track model progress. 4. **Data Visualization:** Use *Tableau* or *Power BI* for creating data visualizations, or *Matplotlib* in Python. 5. **Natural Language Processing (NLP):** *NLTK* and *spaCy* to analyze and understand accident report text.
Hardware:	1. **Computing Power:** High-performance computer with 16 GB RAM, multi-core CPU, and GPU (like NVIDIA) for model training. Cloud options (AWS, Google Cloud) to handle bigger data or run models faster. 2. **Storage:** 1 TB SSD storage for quick data access, plus cloud storage (AWS S3 or Google Cloud Storage) for backup and large datasets

Technology/Technique/ Methodology:	1. **Data Collection & Preprocessing:** Techniques to clean, format, and organize accident data from different sources 2. **Machine Learning Models:** Models like *Random Forest* and *Support Vector Machine (SVM)* to identify accident causes. *Clustering* techniques to find hidden patterns in the data. 3. **Text Analysis (NLP):** Use text analysis to extract important information from accident reports 4. **Classification and Ranking:** Rank accident causes and safety actions based on their impact and frequency 5. **Inherent Safety Design (ISD):** Apply ISD principles, like simplifying processes and reducing risk, for safer designs. 6. **Statistical and Predictive Analysis:** Use techniques like *regression* and *time-series analysis* to predict accident trends.
Algorithm:	1. **Data Analysis:** Use *Decision Trees* and *Random Forest* to categorize accident causes. 2. **Text Analysis:** Apply *TF-IDF* or *Word2Vec* to understand the main points in accident reports. *LDA* for finding themes in the data. 3. **Predictive Models:** Use time-series models like *ARIMA* to predict future accident trends.

Type of Project (Focusing on Data Science):

[√] Data Preparation and Modeling

[√] Data Analysis and Visualization

[ ] Business Intelligence and Analytics

[√] Machine Learning and Prediction

[ ] Data Science Application in Business Domain

Status of Project:

[√] New

[ ] Continued

If continued, what is the previous title?

SECTION C: Declaration

I declare that this project is proposed by:

[√] Myself

[ ] Supervisor/Industry Advisor ( )

Student Name: Haslinda binti Abdul Sahak

Signature **Date** 15/11/2024

SECTION D: Supervisor Acknowledgement

The Supervisor(s) shall complete this section.

I/We agree to become the supervisor(s) for this student under aforesaid proposed title.

Name of Supervisor 1: ....

....
Signature **Date**

Name of Supervisor 2 (if any): ....

....
Signature **Date**

SECTION E: Evaluation Panel Approval

The Evaluator(s) shall complete this section.

Result:

[ ] FULL APPROVAL [ ] CONDITIONAL APPROVAL (Major)*
[ ] CONDITIONAL APPROVAL (Minor) [ ] FAIL*

* Student has to submit new proposal form considering the evaluators' comments.

Comments:

Name of Evaluator 1:

Signature **Date**

Name of Evaluator 2:

Signature **Date**