Vaasan yliopisto
UNIVERSITY OF VAASA

Juha-Matti Toivainen

A. I. Utilization in the Construction Business

A review on present state and potential for Elenia Oy

School of Technology and Innovations
Master’s thesis in Smart Energy
Master of science in Technology

Vaasa 2023

VAASAN YLIOPISTO
Tekniikan ja innovaatiojohtamisen yksikkö

Tekijä:	Juha-Matti Toivainen
Tutkielman nimi:	A. I. Utilization in the Construction Business : A review on present state and potential for Elenia Oy
Tutkinto:	Diplomi-insinööri
Oppiaine:	Smart Energy, Master of science in technology
Työn valvojat:	Petri Välisuo, Hannu Laaksonen
Työn ohjaaja:	Joonas Tutti
Valmistumisvuosi:	2023 **Sivumäärä:** 71

TIIVISTELMÄ:

Tässä diplomityössä selvitettiin rakentamisliiketoimintoihin liittyviä tekoälyn käyttökohteita. Nykyisin liiketoiminnoissa keskitytään operatiivisten toimintojen turvallisuuteen. Projektiliiketoiminnassa projektien ja portfolioiden johtaminen yhdessä turvallisuusjohtamisen kanssa on huomattavan tärkeää. Tiedon puute on harvoin juurisyy ei-toivotuille poikkeamille. Useammin poikkeamat prosesseissa johtuvat epäsäännöllisyydestä ohjeistuksien ja sääntöjen noudattamisen suhteen. Tekoälyyn pohjautuvien työkalujen, kuten koneoppiminen, avulla on mahdollista kehittää turvallisuuteen ja projektijohtamiseen liittyvien tehtävien tehokkuutta. Tutkielma sisältää yleisen katsauksen tekoälyyn ja tarkastelun nykyisistä lähestymistavoista tekoälyn hyödyntämiseen rakentamisliiketoimintoihin liittyen. Lisäksi työssä muodostetaan ehdotukset tuleville vaiheille tekoälyn hyödyntämiseen Elenian rakentamisliiketoiminnassa. Ensimmäisessä osassa käydään läpi yleiskatsaus tekoälyyn liittyen. Toisessa ja kolmannessa osassa työtä tarkastellaan nykyisiä tekoälyn käyttökohteita. Toisessa osassa tarkastellaan rakentamistöiden turvallisuuteen liittyviä hyödyntämiskohteita. Kolmannessa osassa vastaava tarkastelu keskittyy projekti ja portfoliojohtamisen toimintaympäristöön. Yleisin tapa hyödyntää tekoälyä on selvittää ja tunnistaa toimintaympäristön riskeihin liittyvien tekijöiden suhteita toisiinsa. Erilaisissa toimintaympäristöissä on erilaisia riskejä, joiden esiintymisen todennäköisyyttä on syytä pienentää. Koneoppimismallien rakentamisen toteutus on käyttökohde sidonnainen, joten on monia tapoja hyödyntää koneoppimista. Elenia Oy:n toiminnassa projektit ja niiden hallinta ovat keskeisessä osassa mahdollistamassa yhtiön missiota: Elämää sähköistämässä. Sähköverkot vaativat jatkuvaa kunnossapitoa ja johdonmukaista kehittämistä. Osa tätä kehittämistä on teknisen käyttöiän saavuttaneiden komponenttien uusinta, esimerkiksi Elenian Säävarma-hankkeissa. Työturvallisuuden edistämiseksi Elenia on yhdessä kumppaniensa kanssa allekirjoittanut Turvallisuusmanifestin, jonka keskeinen teema on mahdollistaa kaikkien Elenian töissä olevien henkilöiden turvallisen palaamisen terveenä kotiin. Tutkielman keskeisenä lähestymistapana oli etsiä laajasti erilaisia tapoja hyödyntää tekoälyä liittyen turvallisuus- ja projektitavoitteiden kehittämiseen.

AVAINSANAT: Project management, Safety management, Safety, Artificial intelligence, Machine learning, ML, AI, Portfolio management, Construction, Construction business, Risk management, DSO, Distribution Network

UNIVERSITY OF VAASA
School of technology and innovations
Author: Juha-Matti Toivainen

Title of the Thesis: A. I. Utilization in the Construction Business : A review on present state and potential for Elenia Oy

Degree: Master of Science in Technology
Programme: Master's Programme in Smart Energy
Supervisors: Petri Välisuo, Hannu Laaksonen
Instructor: Joonas Tutti
Year: 2023 **Sivumäärä:** 71

ABSTRACT:

The thesis examines the present applications of artificial intelligence in the construction business domain. Nowadays, businesses are focusing on the safety of an operating environment. In a project-based business, managing projects and portfolios with safety management is significantly important. Lack of knowledge is rarely a root cause of undesired deviations. More often, the deviations in processes are related to an irregularity in compliance with the instructions and rules. With the assistance of AI-based tools, such as machine learning, one can improve efficiency on safety and project management tasks. The thesis provides a general view of artificial intelligence and a review of present approaches on AI utilization in the construction domain. Also, the thesis suggests the next steps for the utilization of AI in Elenia's construction business. The first section of the thesis gives an overall view of artificial intelligence. In the second and third sections, a review of the present utilization approaches is examined. In the second section, the utilization is examined in the construction site safety domain. In the third section the examined field is related to the project management domain. The most common way to utilize AI were to exploit existing data for risk prediction and relationship detection. The risks differ from the examined domain. Thus, building a machine learning model is use-case related. There are various ways to utilize different models to achieve the benefits of machine learning. In Elenia Oy's activities managing projects have a key role for achieving company's mission: Electrifying life. The electric grids demand continuous maintenance and consistent development. One part of the development is replacement of components that have reached end of the technical lifecycle. For example, replacement can be executed in Elenia's Säävarma projects. The development of occupational safety Elenia together with its partners has committed for safety manifesto. The key theme of safety manifesto is to render everyone related to Elenia's work field to return home in good health. The key approach of thesis was to find widely different approaches to utilize an AI for the development of safety and project objectives.

KEYWORDS: Project management, Safety management, Safety, Artificial intelligence, Machine learning, ML, AI, Portfolio management, Construction, Construction business, Risk management, DSO, Distribution Network

Contents

LIST OF FIGURES

Abbreviations

AI Artificial intelligence

ANN Artificial neural network

BIM Building information model

CEM Construction engineering and management

CNN Convolutional neural networks

DNN Deep neural network

EPC Engineering, procurement, and construction

GA Genetic algorithm

GDP Gross domestic production

HSEQ Health, safety, environment, and quality

IoT Internet of things

NLP Natural language processing

NLTK Natural language toolkit

NSE Nash-Sutcliffe efficiency

MAE Mean absolute error

MAPE Mean absolute percentage error

ML Machine learning

RF Random forest

RMSE Root means squared error

R^2 Determination coefficient

SVM Support vector machine

1 Introduction

What if the data could describe what happened, why did it happen and what will happen next? Could one use this kind of ability to develop an enterprise's, or individual's, actions for achieving better results for the desired objectives?

In present day the enterprises have comprehensive opportunities to use data for developmental purposes because most of management work is done through system that collects and stores data in frequent basis. An existing data often reflects priorities of the enterprise. Hence, the data have often key performance indicators stored into it, the data can explain the history of operative actions. Also, the outcomes are often determined by these key attributes. The data is, therefore, a backbone of utilizing an artificial intelligence with machine learning applications.

The machine learning models can be used for various analysis purposes: descriptive, diagnostic, predictive and prescriptive analysis. The descriptive analysis describes what has happened and diagnostic analysis answers the question why it happened. Furthermore, predictive analysis aims to describe what will happen in future. Hence, the predictive analysis is not able to answer what must be changed for achieving better results there is demand for a prescriptive analysis. A prescriptive analysis aims to describe what part of the process should be changed for achieving development.

The Thesis' aims to conclude the present state of AI utilization in construction related business. After compiling the present state of AI utilization in the construction business domain for a theory base, the thesis concentrates on leading this knowledge to solutions that enterprise needs to develop project portfolio management and safety in construction sites. Thus, Elenia's operating domain concentrates on project management a literature review's focus is on the projects and safety management. Latter is integrated fixedly into the work life. No matter what enterprises' operational domain is, the development of the safety management, and therefore an overall safety of the operations, is essential for every enterprise. The literature review’s use cases focus on the machine learning solutions. Hence, the objective for the Thesis was to establish present utilizations of the AI in the construction related project and safety management domain a structure and content of the Thesis relates mainly to machine learning applications. Furthermore, it is noteworthy that machine learning is not mandatory for utilizing the AI. The intelligence of the machine may also be beneficial without learning dimension. For example, a chatbot solutions can utilize the AI without the machine learning solutions, chatbots can operate with predetermined rules that the bot follows.

AI based systems often need a large quantity of a data for work properly. This data is often referred to as big data. The big data structure and management could be a subject of a thesis work itself, but the big data and its sub-subjects are discussed only negligibly in this thesis. AI systems uses the data for learning purposes and after learning AI can serve us in business-related tasks. Also, thus AI solutions are saving scarce time of the experts for more essential tasks.

2 What is Artificial Intelligence

Artificial Intelligence is broad concept that can be divided into multiple subtopics, see Figure 1. Artificial intelligence can be described in a one sentence as for example: A machine being able to perform intelligent actions. There are also many other ways to describe AI, hence the concept being overwhelmingly wide and complex. Nevertheless, according to Russel & Norvig (2014) the definition of the AI could be divided into human like and rational acts. The former aims to mimic human behavior in thinking and acting whilst the latter is based on predetermined rules for making rational decisions.

In *Roles of artificial intelligence in construction engineering and management: A critical review and future trends* Pan & Zhang (2021) argues that artificial intelligence deals complex and dynamic decisions with a better accuracy compared to old systems. The paper examined publications considering the artificial intelligence adoption in the construction engineering and management (CEM). They concluded that there is exponential growth of the papers related to AI and its present trends. The authors also stated that an AI acts as a backbone for the future digitalization processes. Hence, the construction industry includes project-based businesses AI solutions are connected tightly to projects.

There are multiple different solutions to mimic advantageous behavior by a machine. It is noteworthy, that human behavior is not considered to be a goal itself because human action may be inconsistent quite often. The logical actions of human are generally target of an artificial intelligence. Computer vision utilizes the machine for seeing, speech recognition makes a machine able to hear and machine learning models, gives ability for a machine to process and use an information collected.

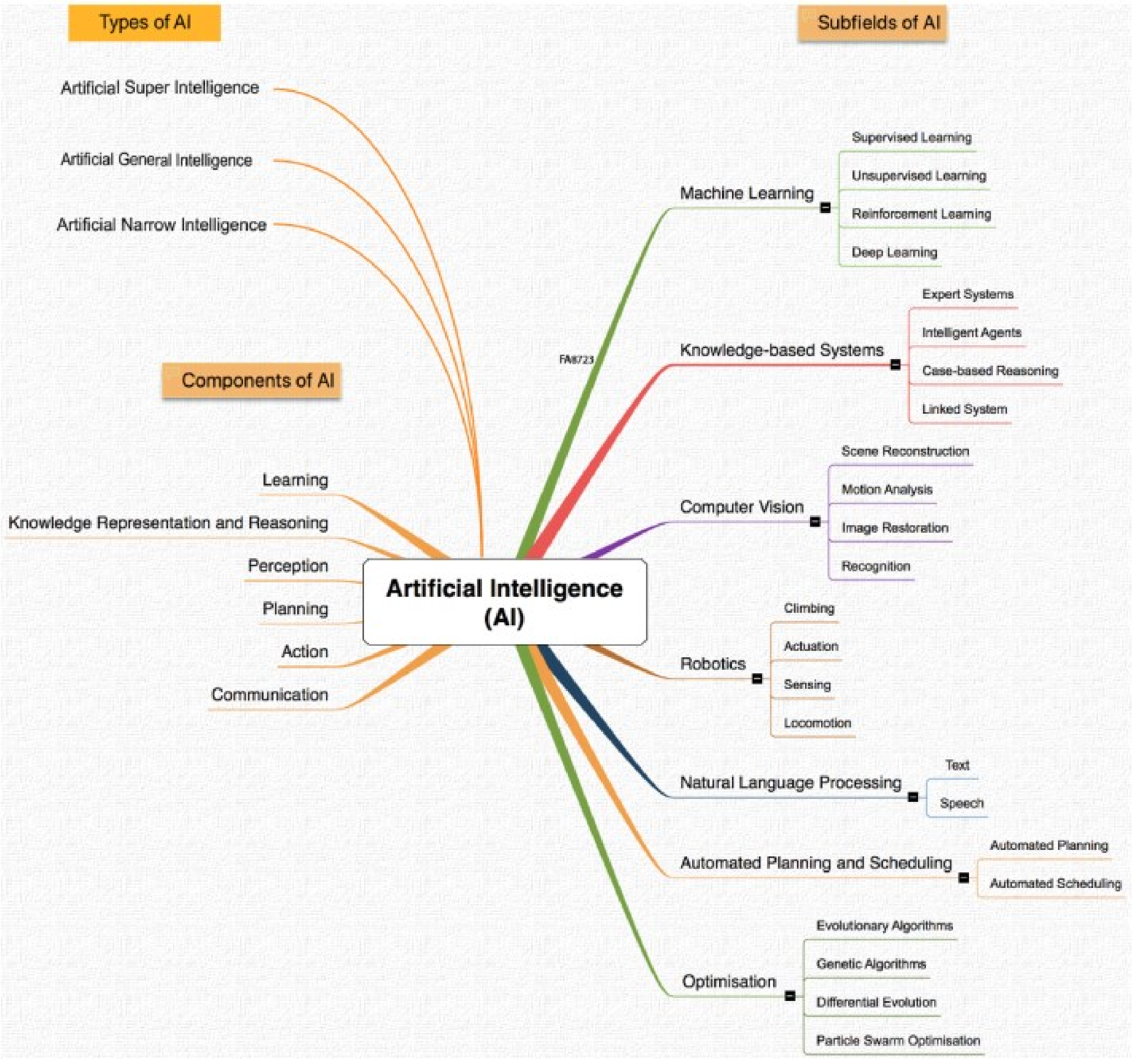


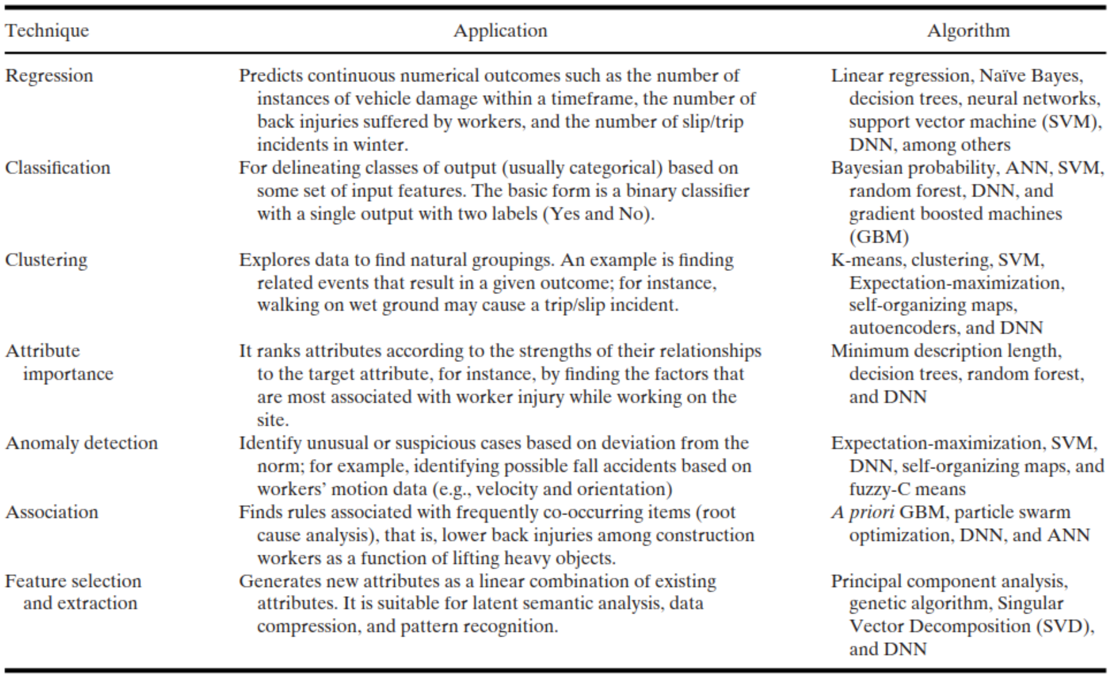
Figure 1. Illustration of artificial intelligence's concept (Abioye et al., 2021)

In Figure 1 the concept of the AI is illustrated with subfields of AI. Left side of the figure describes types of AI. Abioye et al. (2021) explains that the artificial narrow intelligence is type of AI that operates in narrow predetermined domain, for example, repetitive sales prediction. The artificial general intelligence refers to human like behavior by the machine. In this form AI has general learning abilities and the ability to solve problems without predetermined rules, such as humans can learn by examining prevailing circumstances. The artificial super intelligence is the form of AI where machines abilities surpass the human's capabilities in multiple domains. Components of the AI is illustrating the actions that one is keen to achieve with the machine's intelligence. Subfields of the AI are describing the methods and mode of action to achieve the actions.

In *Artificial intelligence: 101 things you must know today about our future* the author discusses data being more valuable than an oil at the beginning of the industrial era. The author argues that an oil was beneficial for handful of companies, but a data is beneficial for much larger quantities. The author also discusses that artificial intelligence is fourth industrial revolution.

3 Machine Learning

Objective of the machine learning is to teach a machine to act with advantageous matter. Often, this action is exploring data and generating predictions from the gathered information. Predicting is done with a model that has learned to predict variables from a training data set. Datasets are divided into learning, validation, and testing datasets; former is used to teach a model to make prediction, validation is used to tune a model for better accuracy and latter is used to test the model. Test-set must always be unseen data for the model. In this section utilized construction business related machine learning tools are discussed.



Technique	Application	Algorithm
Regression	Predicts continuous numerical outcomes such as the number of instances of vehicle damage within a timeframe, the number of back injuries suffered by workers, and the number of slip/trip incidents in winter.	Linear regression, Naïve Bayes, decision trees, neural networks, support vector machine (SVM), DNN, among others
Classification	For delineating classes of output (usually categorical) based on some set of input features. The basic form is a binary classifier with a single output with two labels (Yes and No).	Bayesian probability, ANN, SVM, random forest, DNN, and gradient boosted machines (GBM)
Clustering	Explores data to find natural groupings. An example is finding related events that result in a given outcome; for instance, walking on wet ground may cause a trip/slip incident.	K-means, clustering, SVM, Expectation-maximization, self-organizing maps, autoencoders, and DNN
Attribute importance	It ranks attributes according to the strengths of their relationships to the target attribute, for instance, by finding the factors that are most associated with worker injury while working on the site.	Minimum description length, decision trees, random forest, and DNN
Anomaly detection	Identify unusual or suspicious cases based on deviation from the norm; for example, identifying possible fall accidents based on workers' motion data (e.g., velocity and orientation)	Expectation-maximization, SVM, DNN, self-organizing maps, and fuzzy-C means
Association	Finds rules associated with frequently co-occurring items (root cause analysis), that is, lower back injuries among construction workers as a function of lifting heavy objects.	*A priori* GBM, particle swarm optimization, DNN, and ANN
Feature selection and extraction	Generates new attributes as a linear combination of existing attributes. It is suitable for latent semantic analysis, data compression, and pattern recognition.	Principal component analysis, genetic algorithm, Singular Vector Decomposition (SVD), and DNN

Figure 2. Overview of the machine learning techniques and algorithms (Ajayi et al., 2020).

Ajay et al. 2020 gathered main machine learning techniques into the table, see Figure 2. The table might be useful when one determines the need for the machine learning model. The authors illustrate the application from the perspective of safety domain.

There are four main orientations of machine learning; supervised machine learning, unsupervised machine learning, reinforcement machine learning and deep learning (Abioye et al., 2021). The difference between these disciplines is the state of human interaction in a learning process. In supervised machine learning, the user's interaction is greatest, whilst in deep learning, the machine is making decisions quite self-referential. Thus, the user's interaction is lowest in deep learning. In unsupervised learning the machine learns more independently from the data. The main task of unsupervised learning is to generate relevant information from non-structured data (Pan & Zhang, 2021).

3.1 Artificial Neural Network

The Artificial neural network (ANN) is designed to mimic human brain operating principle. ANN has multiple nodes that forward and receive information with non-linear basis. The system contains visible and hidden layers of nodes (Lin et al., 2021). Also, Lin et al. (2021) describes that ANN can perform the predictions without being taught specific relationships between the data's attributes because the information flows from different input nodes to multiple hidden nodes. The model consists of neurons the information flows in various directions. Hence, the network can generalize learned information. Thus, the model has decreased dependency on the single attribute's effect on the prediction.

Koc et al. (2022) explains that ANN mitigates two-way calculations made in the network, feed forward and back propagation. First mentioned generates random weights from the input data and latter solves optimal weights by using different parts of the neural network.

3.2 Deep Learning

In the article *Deep Learning* LeCun et al. (2015) explained that conventional machine learning models has limitations to handle a raw data from the nature. Deep learning brought development for this domain. For example, in the computer vision-based recognition tasks the first, second and third layers detects different parts of the picture. In other words, different layers detect different features from the data. This allows models develop during the process. Hence, there is multiple layers that can individually learn different parts of the picture, and later gather the knowledge, the learning becomes deeper compared to single layer models.

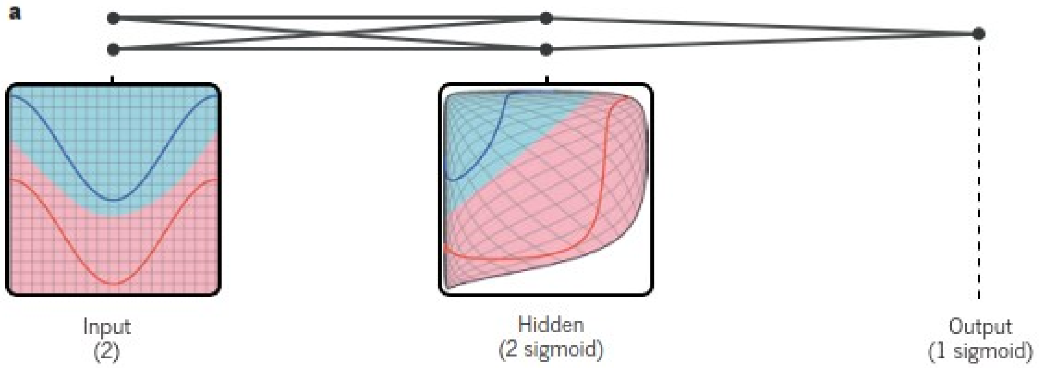


Figure 3. Illustration of a multi-layer network classification of a data (Lecun et al, 2015).

Lecun et al. (2015) illustrated, see Figure 3, how multi-layer neural network can bend the data space to make a data linear. Red and blue lines represent a categorical data, area of the red and blue data is non-linear at the left side of the figure and the hidden layer process the regular grid for the linear mode. The authors discuss that deep learning related subjects such as Backpropagation and feedforward architectures. Backpropagation is the concept where the deep learning models is calculating different weights for the layers, the weights affect to the prediction that model is producing. Feedforward structures calculate these weights during the process while moving between the layers. The model is optimizing the weights for best prediction accuracy (LeCun et al., 2015).

3.3 Support Vector Machine

Koc et al. (2022) describes Support Vector Machine (SVM) as the multiuse model due it universal structural learning process. The SVM generates a hyperplane and optimizes it regarding to the datapoints. Thus, the hyperplane represents the optimal mean regarding to the datapoints there is error with the non-linear data. This demands one to use "kernel tricks".

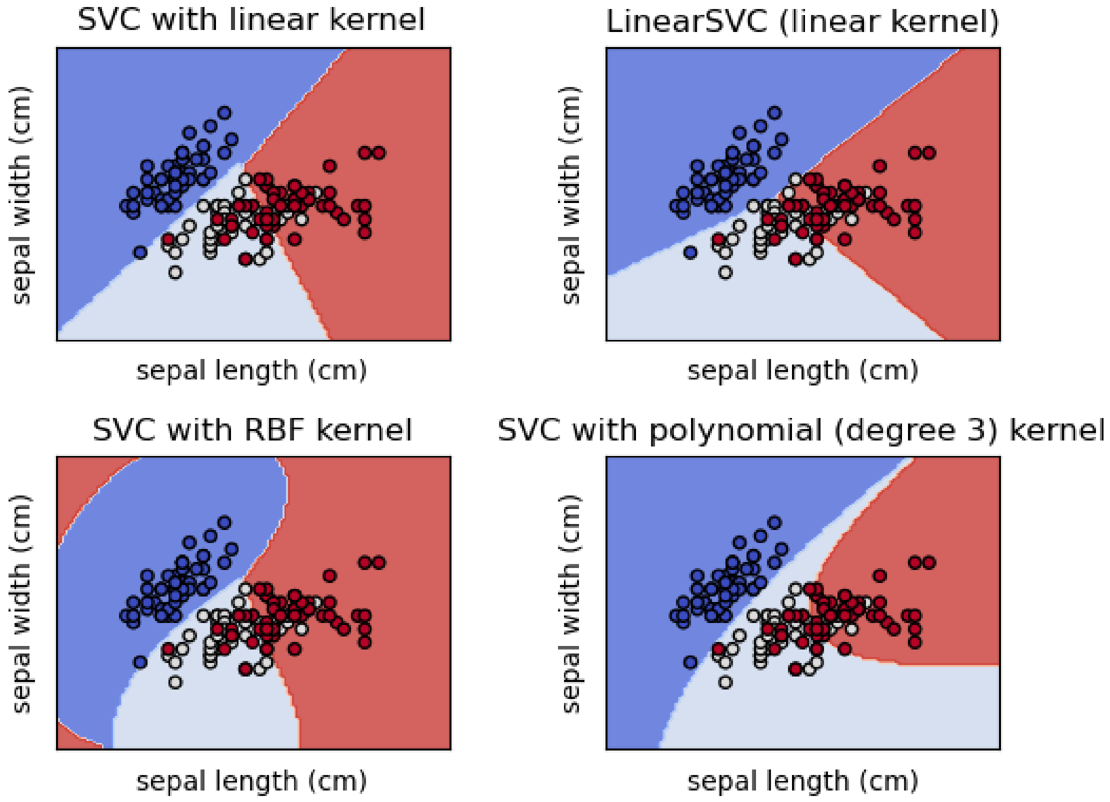


Figure 4. Illustration of four different kernels used in SVM (Pedregosa et al. (2011).

Lei (2016) explains that by applying kernel function the sample data is placed in high-dimensional map where the non-linear classification becomes possible. By generating high dimensions for the datapoint mapping the support vector can gain higher resolution when obtaining support vectors. Hence, the decision boundary follows data in higher resolution it is not linear, see Figure 4. Note that there are examples of two linear and two non-linear kernels. Also, by using kernel functions it is possible to generate linear decision boundary if the data points are taken from the two-dimensional map to the three-dimensional map. Non-linearity can disappear by adding the y-axis.

3.4 Natural Language Processing

Natural language processing is used to identify descriptive key words or phrases from the written text. Cheng et al. (2020) explains that the Natural Language Toolkit (NLTK) is most popular library for NLP. The toolkit is providing multiple different components for text modification such as a tokenization, stemming and parsing, for example. With these a user can make natural written language to more suitable for the machine learning (ML) model to use it for prediction.

In *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, SpaCy, and Keras* Srinivasa-Desikan (2018) explains that Gensim is a Python library dedicated to text processing, particularly in vectorizing the text. Also, the author notes that abilities of the Gensim extend further. Vectorizing the text is beneficial for the ML prediction making process. After vectorizing a word, it can be represented in a mathematical form. The vector has a magnitude and direction so it can be represented in the multiple dimensions. The author explains that Gensim is memory independent, and it can use multiple implementations of semantic domain. Also, smooth operation in the Python's ecosystem helps to co-operate with multiple vectorizing tools and algorithms.

3.5 Convolutional Neural Networks

LeCun, Bengio and Hinton (2015) in *Deep Learning* discusses that convolutional neural networks (CNN) principle is to detect local combinations of the features and a pooling layer connects these local results together. Architecture of a CNN divide features in many layers, content of these layers is divided to feature maps. Layers are connected to feature map and filter banks. The model is feedforwarding the weights to next layer.

3.6 Recurrent neural networks

LeCun, Bengio and Hinton (2015) explained that Recurrent neural networks (RNN) are powerful tools when input data have sequential inputs. The network processes the data in the hidden units and creates a state vector that got a history of the process in it. The authors note that RNN's training can cause problems due the backpropagation gradients, these gradients tend to over- or underfit the prediction. Thus, the gradient tends to be similar to the earlier gradient.

3.7 Decision trees

Pedregosa et al. (2011) explain that decision trees are used for the classification and regression problems. The algorithm exploits simple decision rules during the learning process. The decision nodes are generating the tree structure of the model. The authors note that decision trees have tendency to overfit if the data have many features. Thus, it is important to decrease the probability of overfit by suitable method, such as maximum depth of the tree. The limitation in the tree depth decreases the amount of vertical decision nodes, therefore, the model is not able to exploit all information of the data's features. Thus, the model's ability to perform predictions with unseen data is better.

Pedregosa et al. (2011) explain that because the decision tree's prediction is not continuous either smooth the variance of the individual tree's prediction can be quite high. By using the random forest individual decision tree's prediction weight is decreasing and therefore the variance of the prediction is lower. Lin et al. (2021) explains that Random Forest (RF) is suitable for a nonlinear regression and a classification machine learning problem. Therefor RF is popular in the risk prediction ML domain due its ability to solve nonlinear relationships of variables. Random forest works in equivalent manner that decision tree. Structure of the decision tree is a seed for the random forest. In random forest the algorithm is producing multiple trees that are used as an ensemble for making the prediction.

3.8 Optimization

Pan & Zhang (2021) discuss the optimization of a project. In project management domain the optimization can be considered an optimal decision support system to achieve the best end-result for a project.

In the machine learning domain, the purpose of the optimization is to decrease the prediction error. Hence, the model's ability to generate more precise prediction increases. At the process of the building a model for prediction one should divide the data into training, validation, and testing datasets. By this simple procedure one can evaluate and validate the model's performance on the unseen data. The model's performance is measured multiple times during the process, see section 1.2.3 Performance metrics of a model.

In *Machine Learning: An Applied Econometric Approach* Mullainathan & Spiess (2017) explains that the performance of the model may be overrated in the training data. The authors explain that some algorithms tend to overfit. Overfit is a situation where the model learns training attributes of the data too well. Thus, the model can perform significantly better on the training set compared to testing or validation dataset where data attributes deviate from learned ones. The authors explain that part of the solution is regularization that measures the complexity of the model and directs the optimization towards simpler models. For example, in the regression tree-based model one can perform regularization by comparing the tree depth and performance. Thus, one may not select merely the best overall performing model because the regularization demands to also take the model complexity in concern, such as tree depth.

3.8.1 Performance metrics of a model

A model's accuracy must be validated with some objective practice. It is intuitive to measure a model's performance with multiple indicators. The evaluation metrics measure the error between the prediction to the actual value. At the classification the metric measures model’s ability to classify data inputs into correct categories. In the regression problem the evaluation is based on the numerical value between the actual data point and the predicted one.

Oyedele et al. (2021) describes the evaluation metrics represented above shortly and accurately:

> *“Precision is a fraction of correct predictions for a specific class, while recall is the model’s ability to classify relevant cases. F-1 score defines the harmonic mean (or a weighted average) of precision and recall, and it reaches its best value at one and its worst at zero.”*

Koc et al. (2022) used five different indicators for the regression-based machine learning problem: Mean Absolute Error (MAE), Root Means Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Determination Coefficient R^2 and Nash-Sutcliffe efficiency (NSE). The authors explained that first three indicators represent higher accuracy with smaller absolute values. The determination coefficient near -1 and 1, thus zero represents non-acceptable prediction accuracy. NSE prediction accuracy moves from 0,5 to 1 where 1 is perfect prediction.

Oyedele et al. (2021) discusses of the model’s performance analytic tool Cohen Kappa in their paper. Cohen Kappa is suitable for the classification-based problem’s performance evaluation. The authors discuss that Cohen Kappa fits in the multi-class and biased problems measurement. The equation for the Cohen Cappa is described in equation (1).

$$k = \frac{(t-y)}{(1-y)} \tag{1}$$

Where the y is predicted output and t is value of predicted variable, 1 represents the perfect performance of the model.

The authors used multiple other evaluation metrics; Accuracy, Precision, Recall and F1 Score. The evaluation metrics are described in equations (3) – (6) as follows:

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{(TP + TN + FP + FN)} \quad (2)$$

$$Precision = \frac{TP}{TP + FP}$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

3.8.2 Cross Validation

According to Mullainathan & Spiess (2017) cross validation is a tool for selecting correct level of the regularization. In the cross-validation one splits the training data to equal sized sections, often called folds. Thus, one has, for example, ten training dataset over one larger one. One can see from the Figure 5 that data is divided into different folds and splits. The model is tuned by using separate folds as testing data in different splits.

Cross validation is blocking overfitting of the model by separating data to the folds. The model is tested against test set multiple times during the learning process. Thus, the model's performance is tested in every split. Therefore, the feedback from multiple testing events is used for tuning the model's parameters into right direction.

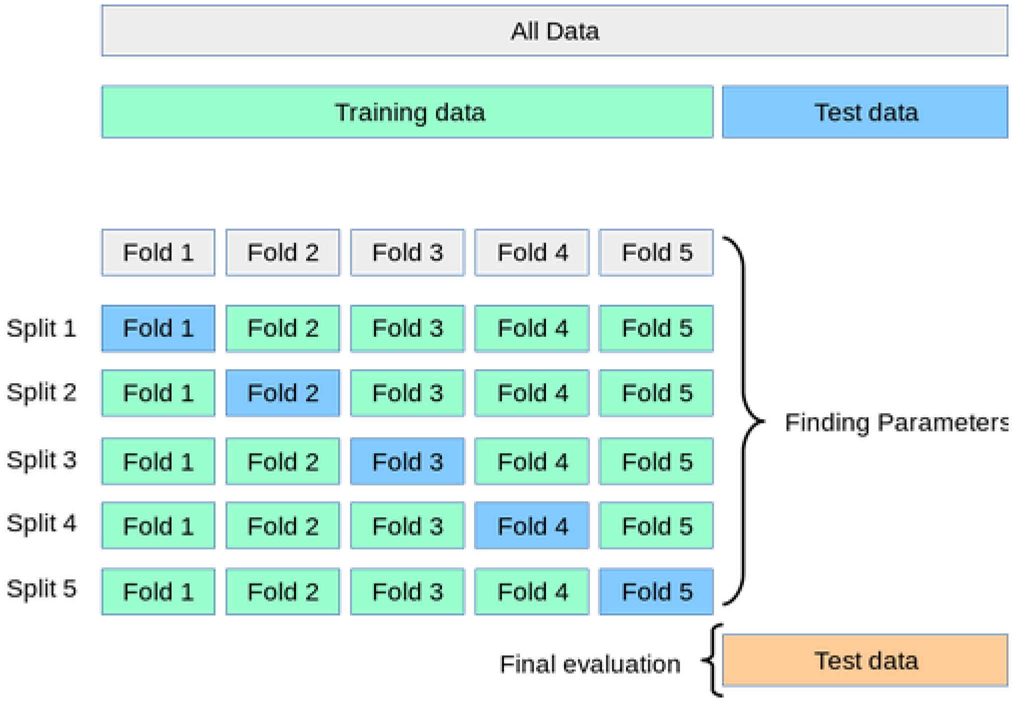


Figure 5. Illustration of the k-fold cross-validation principle (Pedregosa et al. (2011).

In other words, the model has better ability to consider the attributes and relationships that are important for the prediction accuracy in unseen dataset. Also, the advantage of cross validation multiple testing events is that the data can be tested only once. After the data is tested against test set the model has seen the data and the test can't be repeated on the same data set. It is essential to train model for making accurate predictions with unknown data. Therefore, the cross validation is using data with the greater efficiency compared on the conventional data set structure.

3.8.3 Genetic Algorithms

There are multiple ways to optimize the model. Different methods are suitable for different problems. For example, Gradient Descent algorithm calculates gradient and iterates through minimum value of the data. Hence, the gradient for the data can't be calculated in every data set there is need for different algorithms. Some algorithms iterates on the linear basis these can't be used in the non-linear data effectively, also the data shape can limit the algorithms ability to find minimum and maximum values of the data.

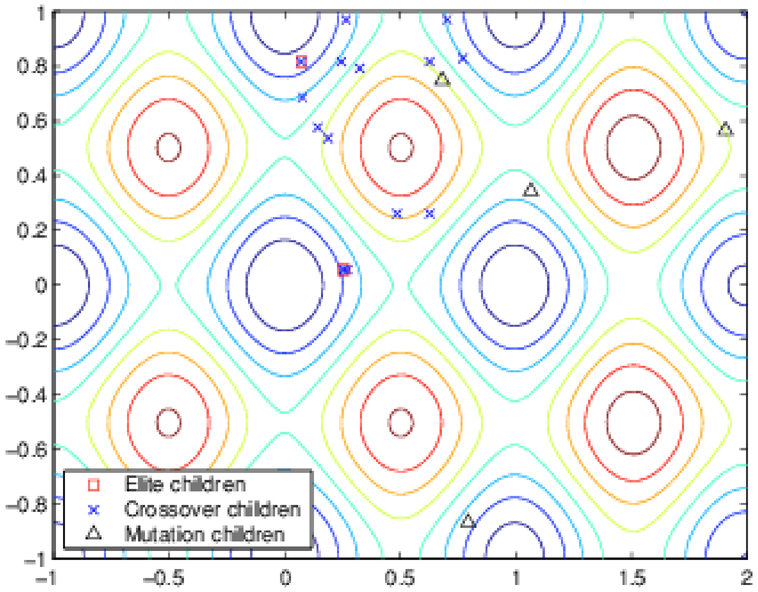


Figure 6. Example of the GA generated random population and explanation of the datapoints in the next generation (MathWorks, n.d.).

Genetic Algorithms (GA) is one way to optimize the model. In the thesis's domain, safety and project management, the GA is popular way to optimize the models. Thus, GA is introduced in this subsection. Skorpil and Oujezsky (2020) explains that Genetic Algorithm (GA) is an optimization algorithm that mimics nature's evolutionary development. The authors note that GA is suitable for many problems but is unsuitable for some problems due time usage of the fitness evaluation. An individual execution of the function demands high amount of time.

In GA domain individual datapoint, often referred to gene, is part of the population. The data points that have best abilities compared on the problem that one is solving are selected by GA during the process. The data points are used as the parents for next iterations population. GA can generate the parents in multiple ways, the practice is often referred to as operator. The genetic operators define how next iterations population is born. For example, single point crossover operator chooses random point that divides the population into segments, after this the populations are compounded. Other common operators are selection and mutation. In the selection data points with suitable features are chosen to next iterative population. In the mutative operator the genes are randomly inserted to population, or the features of the genes are altered for achieving variability in the population.

First the algorithm generates random population of the data. The generated population is illustrated in the Figure 6. The datapoints are marked with different patterns that explains evolution of the data points at the next generation. The elite datapoints are directly selected as a parent in the next generation and crossover datapoints are mixed with other crossover datapoints. The crossover selects two parents for a data point, mutation changes the form of the data and selection delivers the data for the new generation as it is. The mutated children's attributes are changed for the next generation. These generations are reproduced till the stopping parameters, such as max generations or maximum time, are achieved (MathWorks, n.d.).

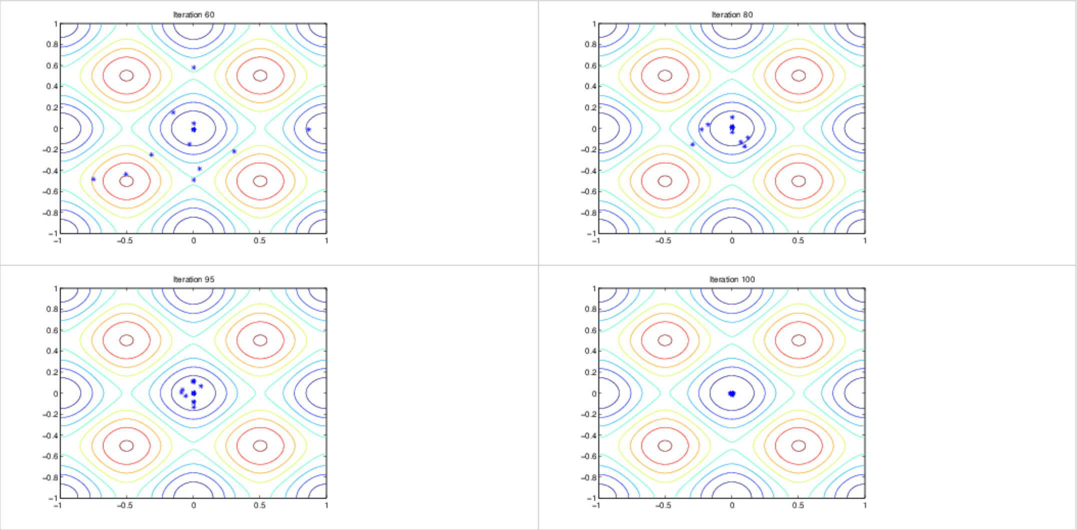


Figure 7. Evolution of the datapoints during the iteration of genetic algorithm (MathWorks, n.d.).

Due the iterations the population is slowly gathering to the center of the diagram, see Figure 7. This is a result of the evolutional nature of the algorithm.

3.9 Coding the machine learning models

Coding of the machine learning models can be done in various ways and the languages has strengths and weaknesses. See below Julia's developers' explanation for creating new language:

> *"We want a language that's open source, with a liberal license. We want the speed of C with the dynamism of Ruby. We want a language that's homoiconic, with true macros like Lisp, but with obvious, familiar mathematical notation like Matlab. We want something as usable for general programming as Python, as easy for statistics as R, as natural for string processing as Perl, as powerful for linear algebra as Matlab, as good at gluing programs together as the shell. Something that is dirt simple to learn, yet keeps the most serious hackers happy. We want it interactive and we want it compiled."* Bezanzon et al. (2012).

3.9.1 Python

A Python is an open-source object-oriented coding language that is often used in the machine learning context. The Python seem to be intuitive and relatively easy to learn. Often machine learning courses suggest Python for a coding platform. The Python is attractive language for the machine learning due it has quite wide package of pre-coded libraries in it. The user can simply use line of code for download the library needed for the current coding work.

In *Learn Python programming: A beginner's guide to learning the fundamentals of Python language to write efficient, high-quality code* Romano (2018) explains that in the programming often one must represent real world connections in the code. In the coding language connections are often represented as objects. Therefore, Python's object-oriented nature is advantageous for practical solutions.

The Python, among many other languages, is used to handle necessary attributes and format of the data. Amount of a data could be massive, and unfortunately all data is not in the same format compared to each other. The enterprise that wants to build a machine learning model usually must process the data in multiple ways. For example, the data could be in the text format and needs to be converted to the numerical variables.

The purpose of the data processing is to clarify structure of the data to make it more understandable. For example, the one can relatively easily handle missing values of the data set by filling thousands of values with few lines of the code. The code can fill the missing values with average of the values or simply delete rows with certain missing values.

3.9.2 Julia

Julia is an open-source programming language. In *Julia Data Science* Storopoli et al. (2021) note that Julia is fast and easy to learn language. The authors claims that Julia is easier to read during debugging compared to Python and R. Also, the authors note that Julia has better ability to adjust on the other languages and open-source packages, thus interfaces are reduced, thus coding is more effective. The authors also noted that the program project management tool and package management.

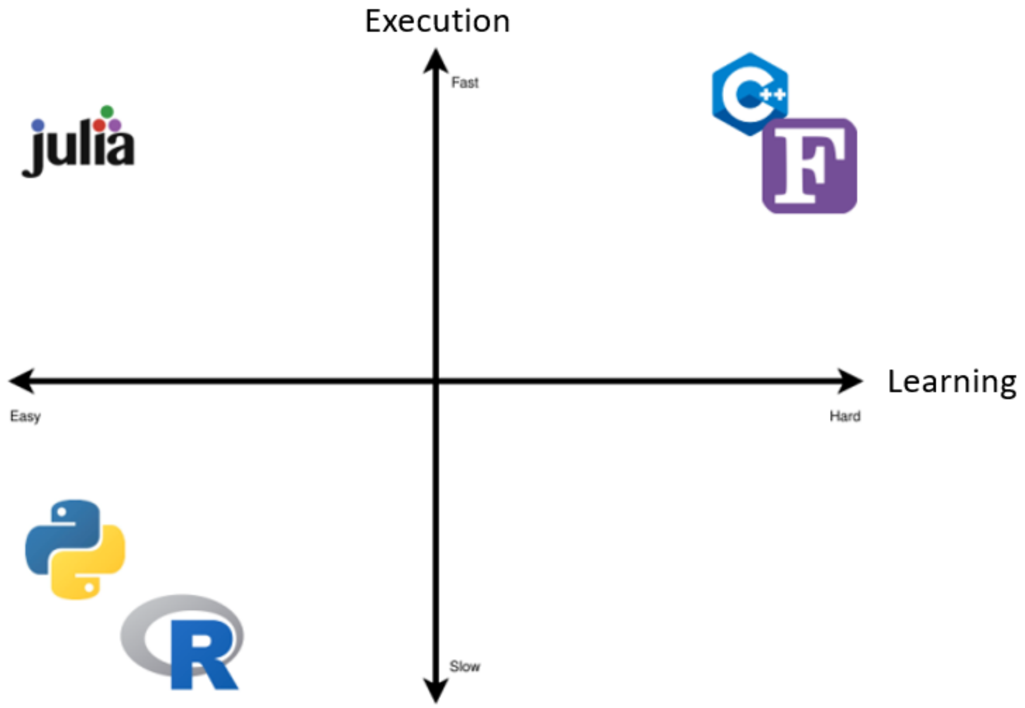


Figure 8. Comparison of the coding languages, regarding to user’s learning rate and code execution (Stropoli et al., 2021).

In Figure 8, the authors classed five open-source languages to quadrants. At the right side of the horizontal axis is languages that are harder to learn and write. At the left side is an opposite language in this manner. The vertical line is dividing languages by the speed of the code execution.

4 Literature review on AI utilization in construction domain

Safety management is a corner stone of a project related to any kind of a construction. Everyone has right to work safely, arrive home from the work. Elenia has committed to the Safety Manifesto. The idea behind manifesto is to connect Elenia and main partners to take responsibility and proactive actions to raise safety levels at Elenia's operating domain. Elenia's main contractor partners has signed the safety manifesto as a commitment to goals of the manifesto.

Project and portfolio management are also key elements for the business to function correctly. The management of projects and portfolios is management of the resources that the enterprise has under control. Hence, the resources are often limited effective management supports resources objectives. An effective use of the scarce resources is therefore beneficial for the enterprise. It is noteworthy, that from the projects perspective the resources are related to projects objective and tasks and does not only comprehend workforce. Therefore, the management of all resources, such as a cashflow, benefits a project to achieve objectives and may determine objectives that can be achieved.

4.1 AI utilization in safety domain

Pan & Zhang (2021) discusses, according to McKinsey global institute (2017), that construction business is responsible for the approximately 13-15% of the world's gross domestic production (GDP) while the construction domain is responsible for the 30-40% of fatal accidents (Zhou et al., 2015). Therefore, a Health, Safety, Environment and Quality (HSEQ) management should be the one top priority for the industry, thus it should be top priority for an individual enterprise and for an individual worker. A worker should be in center of the actions for developing the safety of the construction site. Thus, the construction projects are labor intensive and relatively unique projects it is challenging to see considerable change in near future. Artificial intelligence systems are adopted first in the abstract expert-level and later in the physical activities. Hence, the implementation of the non-physical systems is generally straight forward. For example, AI development in a project management process serves all project resources Thus, using an AI system to physical tasks requires task specific AI system it is costly and time consuming to produce the system for different tasks compared to non-physical systems.

A Natural language processing can be used to process safety reports with relatively low or no cost. Processing the reports can lead prolific knowledge from safety incidents. Baker et al. (2020) used attributes and natural language processing (NLP) to predict injury severity and incident type among other factors. The attributes describe different key elements and conditions of the safety incidents. It is notable that the keywords, attributes, are not outcomes of the safety deviation but the preceding action or circumstance. Also notable is that authors did not use machine learning to extract the keywords while keywords were selected manually. It is notable that manual processing can be boosted by using data processing actions from the machine learning domain.

Also, Baker et al. (2020) constructed machine learning models that predicted safety outcomes. At their work they used the NLP and machine learning modelling. As written above the attributes were extracted so they represented the circumstances before the safety occurrence. They built multiple different models, evaluated models' performance, and stacked the models to achieve best performance. The models can be utilized to identify correlations between safety issues and for example certain tools, the system can be used to identify inverse correlations (Baker;Hallowell;& Tixier, 2020). Thus, the model can be used as a diagnostic inference.

Zhang et al. 2019 examined different machine learning models to classify causes of the accidents. The authors built the model by using sequential quadratic programming to optimize weights for five different models that are represented in Figure 9. At the data preprocessing phase, the accident reports data is structured and cleaned for the future use. The authors executed multiple natural language text processing actions for the accident data. In the model building phase, the authors made five different models for predicting causes of the accidents. In the model tuning phase, the Sequential quadratic programming algorithm optimizes the weights of the models.

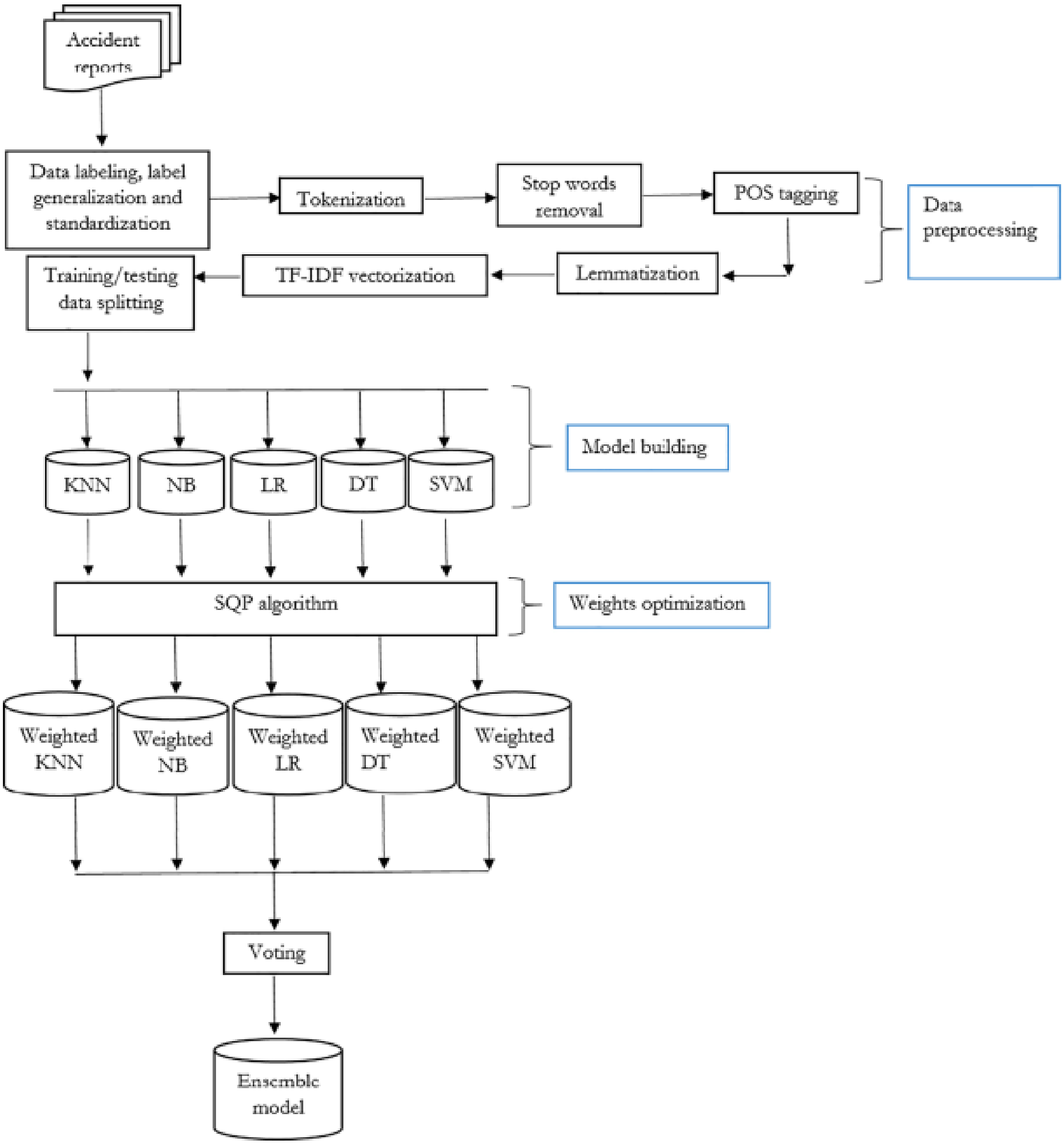


Figure 9. The process of prediction model's development with explanation of the process's phases (Zhang et al., 2019).

The optimized classifiers were able to outperform the classifiers with no optimization. The optimized models F1 score increased significantly. Non-optimized models scored F1 at the range of 0,44 to 0,58 and optimized results were at 0,68. The results describes average F1 score for different accident cause prediction. The author's model predicted 11 different causes for accidents, for example "collapse of object", "Falls" and “electrocution”. It is noteworthy that the optimized model could not predict all the causes with expectable range of F1 score. For the example, the cause of struct by falling object was predicted significantly under 0,5. In other word the model is biased towards wrong direction. With random guessing a model can predict the causes with the 0,5 F1 score. The authors explain the reason behind this is an inaccurate language used at the accident reports. A natural language can have multiple different options to describe similar causes. The authors also claim that in many cases human was not able to extract the correct reason for the accident.

Liang and Liu (2021) examined a safety system with the risk warning and indication control. The authors discussed the system integration with a Building Modelling System (BIM), Internet of things (IoT) and safety risk warning system. The BIM system is used to share information in the building construction process with all the main participants of the project. The authors explain that BIM can be understood as a platform over the conventional design plan.

Also, Liang and Liu (2021) discuss that in the construction building process risk exists among the process all the time. A risk must be measured by its seriousness and probability of occurrence. In a present work field, the variability of the risks is quite high. Hence, the safety culture has developed to a point, where proactive actions block reoccurring incidents. Nowadays a typical risk is reduced by risk mitigation process, this leads decline in typical risk occurrence. Thus, reported incidents and risks are high in variability. The authors discussed to capture three main elements to their early risk warning system. The system had to be able to identify relevant and deviating factors from the construction domain and therefor help a project personnel to mitigate unfavorable probability of a risk with correct actions. The authors approached the systems design from the quality-based context, by focusing to create a measurable and repeatable model that connects science of a safety. An objective approach reduce subjectivity from the safety development process. By reducing subjectivity, an enterprise can increase quality with a measurement among other actions.

At the article *Hazard Analysis: A deep learning and text mining framework for accident prevention* Zhong et al. (2020) discusses similar usage of the system that evaluates hazard reports automatically with a combination of multiple machine learning models and text processing. The authors argue that analyzed repeatable behavior collected from the reports can improve safety. An unorganized reporting due use of a natural language makes limitations to conventional systems to describe data efficiently. In the article, automation was concentrated to exploit the deep learning-based system. By using Deep Learning based models, one can significantly reduce workload of an enterprise's work resources.

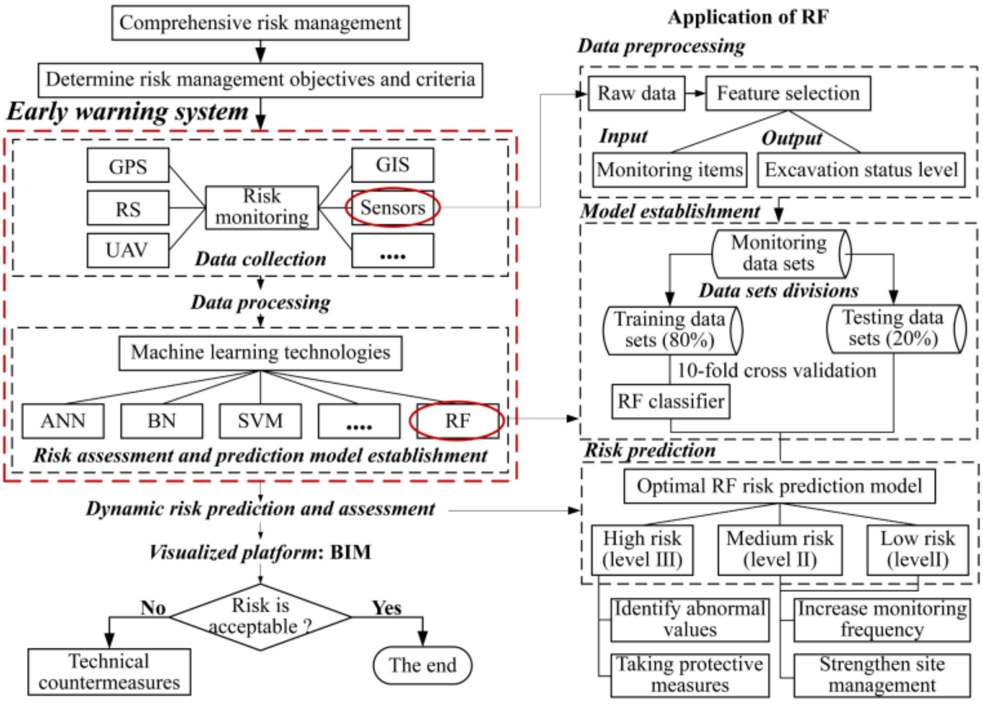
Also, Zhong et al. (2020) analyzed the hazard records with the text mining model to connect relations between different keywords from free-text descriptions. The authors aimed to visualize related causes, common factors, and circumstances of hazards occurred. The model's input feed was gathered from the mobile reporting application that was used in various Wuhan Metro's construction sites. The paper concentrated to the text processing and visualization. Also, there was section where the authors used machine learning model to recognize the relevant keywords. The authors tested the support vector machine and central neuron network for work. The model achieved average F1 score of 0,71, variance of F1 score was quite high.

Goh and Ubeynarayana (2017) examined six different models to recognize safety hazard causes, and factors leading to safety hazards. The authors describe a text mining process that uses tokenization. In natural language text processing tokenization refers to process where the natural language is broken into tokens. The authors used uni-grams and bi-grams tokens that consists of one and two words, respectively. The token is a dynamic expression because the token can be a word, two words or some other specified collection of the words. By making text to numerical values with tokens one is vectorizing the text, thus the machine can better understand the data. The authors describes that their method is not fully automated and proposed modifications to the reports for the better prediction accuracy. The authors explained that misclassification and over-focusing to non-important factors cause challenges to the model's accuracy. To achieve better accuracy the authors, suggest that reporter manually connects cause to pre-set label, for example "electrocution" or "traffic".

Pan & Zhang (2021) discussed ability of an AI in a risk mitigation process. An Artificial Intelligence system can predict risks and work phase's interrelations, therefor helping a project personnel to adapt correct actions at the right time. An Artificial intelligence in risk management process can support an expert with reduction of a subjectivity and an indefiniteness from the safety, or an overall, management process. With an ability to manage multiple information sources at the same time an AI system can produce insights and notifications to a project personnel.

In *Text mining-based construction site accident classification using hybrid supervised machine learning* Cheng et al. (2020) discusses that accident report narratives explain the causes of the occurred accidents. The AI solutions are in active use in present time, but the system's learning ability is limited, and error rate is quite high. A neural network and a recurrent neural network praise these abilities over more traditional models, such as the decision tree, K-nearest neighbors, linear regression, and support vector machine. The authors also discuss about gated recurrent unit that is newer model for a sequential data forecasting.

Lin et al. (2021) describes the early warning system for the excavation works. The authors approached risk assessment and management with a fuzzy set theory together with the machine learning models. The model used the big data together with the sensors to achieve a proactive risk system. The authors identified main risk factors concerning the work domain and derived the knowledge to work phases where risks occurred. The system benefits an analytical hierarchy process and TOPSIS-method, both systems are designed to process multi-criteria decisions. The paper's early warning system starts from the overall risk management which determines the objectives of the risk management process. The system itself consists of two main parts, see Figure 10. The data gathering could be done by multiple ways, such as remote sensing, sensors, geological information system. The data is then processed with random forest algorithm, this is second part of the application where the data is processed. The RF is analyzing the gathered information and producing the excavation status as an output. The outputs are categorized into three risk categories. The predictive actions are chosen by the predicted risk level.



Abbreviations: BIM: building information modelling, RS: remote sensing, GIS: geological information system, GPS: global positioning system, UAV: unmanned aerial vehicle, ANN: artificial neural network, RF: random forest, BN: Bayesian network, and SVM: support vector machine.

Figure 10. The early risk warning system's process by Lin et al. (2021).

Lin et al. (2021) examined the metro subway station as the case building site. The authors gathered and processed a relevant data for the risk prediction system. In the case study data was collected with the sensors buried near the excavation site. The authors seek the system that could predict a risk level to pre-fixed classes. The author's goal was to create the system that indicates risk levels for the construction site personnel to proceed with the relevant actions to mitigate the risks. The authors claims that the system's strength is to predict actual risk level and non-subjectivity of the ML model.

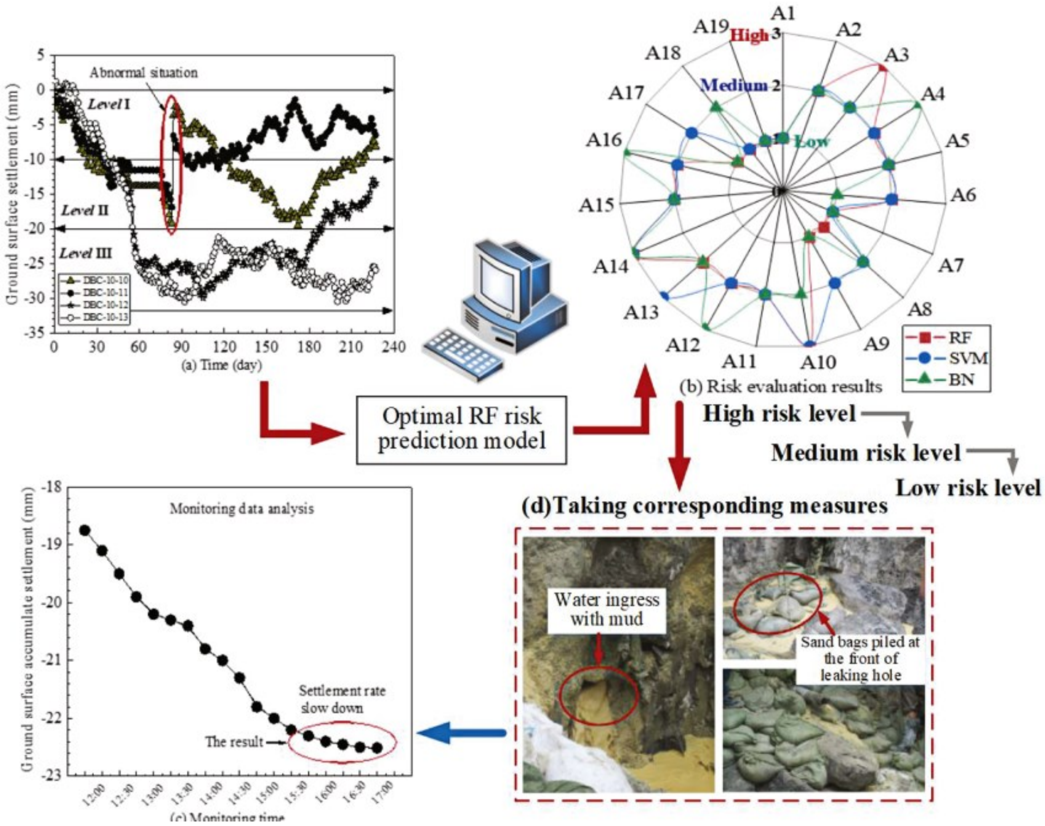


Figure 11. Risk warning system's principle by Lin et al. (2021).

In the Figure 11 authors illustrate the principle of the early risk warning system. The system detects ground surface settlement state from the site's sensors. The risk evaluation is done by random forest, Bayesian network and support vector machine algorithm. Proactive actions are done for lowering the risk levels of the excavation site. Different algorithms are used to decrease the variance of the risk level prediction compared to actual state of the risk. Hence, a model could forecast, for example, low-risk stage during high-risk actual stage it is logical to decrease the variance of the prediction by using multiple models.

At the *Computer vision for anatomical analysis of equipment in civil infrastructure projects: Theorizing the development of regression-based deep neural networks* Arashpour et al. (2022) examined a machine vision together with a deep neural network-based system to raise heavy equipment safety among other beneficial factors. The authors claim that the deep neural network-based machine learning model is accurate in excavation related domain. The vision-based systems utilize multiple layers to determine, for example, the excavator position, see illustrative Figure 12.

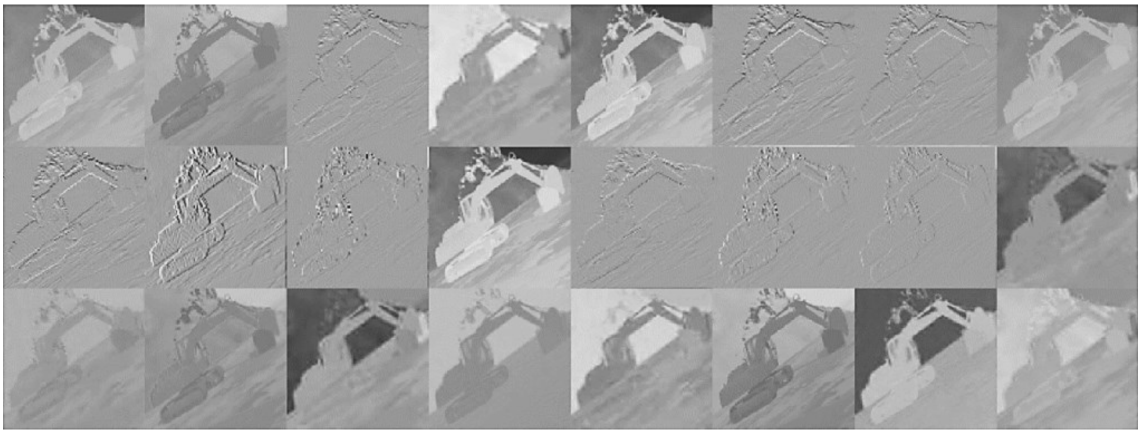


Figure 12. Different layers of the vision-based system (Arashpour et al., 2022).

The system's machine vision focuses on identifying point-form spots of the excavator, such as cabin boom and arm bucket. Also, system aimed to identify the angles of these key points. In the Figure 13 authors are illustrating the data-flow, the system preprocesses the images with channel shuffle, depth wise separable convolution, and compound scaling. After these actions the image's features are easier to detect by the algorithm. Relatively big machinery, such as an excavator, has quite long reachability and demands a lot of space around the working station that makes such machinery possible object to the safety hazards. The preprocessed images are processed by the neural network algorithm. The algorithm is producing a position of the excavation machine as a result of the process.

The authors explain that deep neural network's ability to generate multiple layers is beneficial for the overall efficiency of the system. With a help of the machine vision the excavator operator can control the key points of the machine with a better accuracy leading more precise movements of the machine. The authors note that for the complex objectives the system needs to be trained with a benchmark imaginary data set.

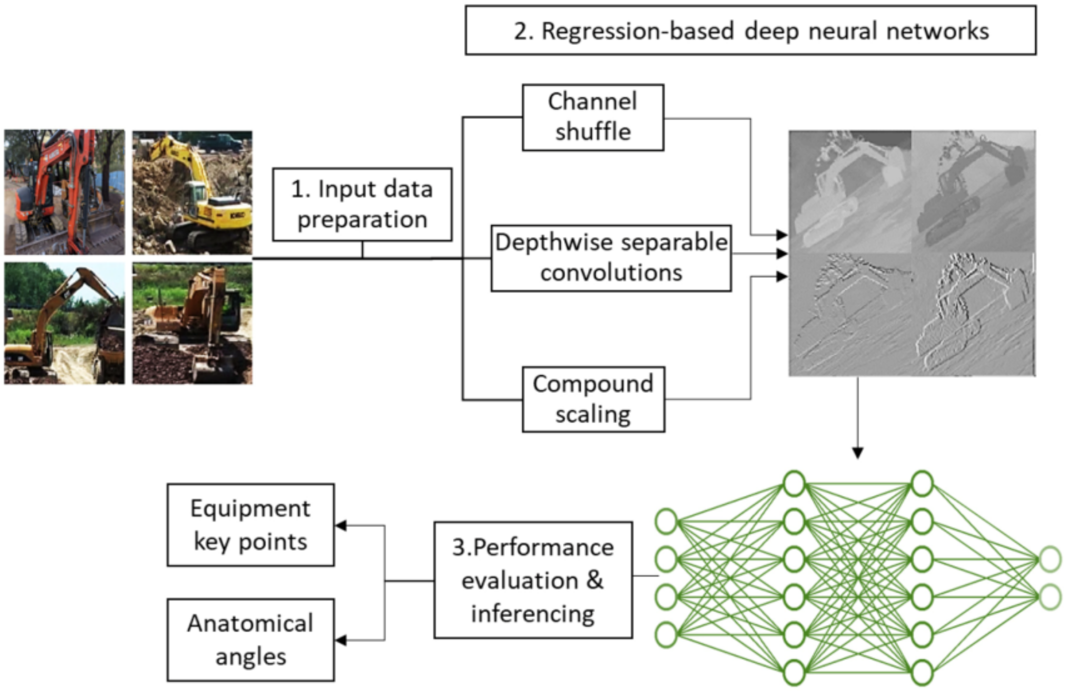

2. Regression-based deep neural networks
1. Input data preparation
Channel shuffle
Depthwise separable convolutions
Compound scaling
Equipment key points
Anatomical angles
3.Performance evaluation & inferencing

Figure 13. The excavation risk warning system's data workflow (Arashpour et al., 2022).

Koc et al. (2022) examined a database of near 400 000 accidents and built a hybrid model with a wavelet and machine learning model. The authors claim that in the present literature there is lack of time series-based approach regarding to the accident prediction. In the wavelet transformation the time series is divided into different bands of the wavelet, where the different bands represent details of the data or approximations with the different wave lengths. The study focused on three different time periods, 1 day ahead, 7 days ahead and 30 days ahead. The authors explained that time series-based prediction is most accurate for the 1 day and 7 days ahead prediction, explanation of this could be day of a week-anomaly, certain days are more dangerous than others.

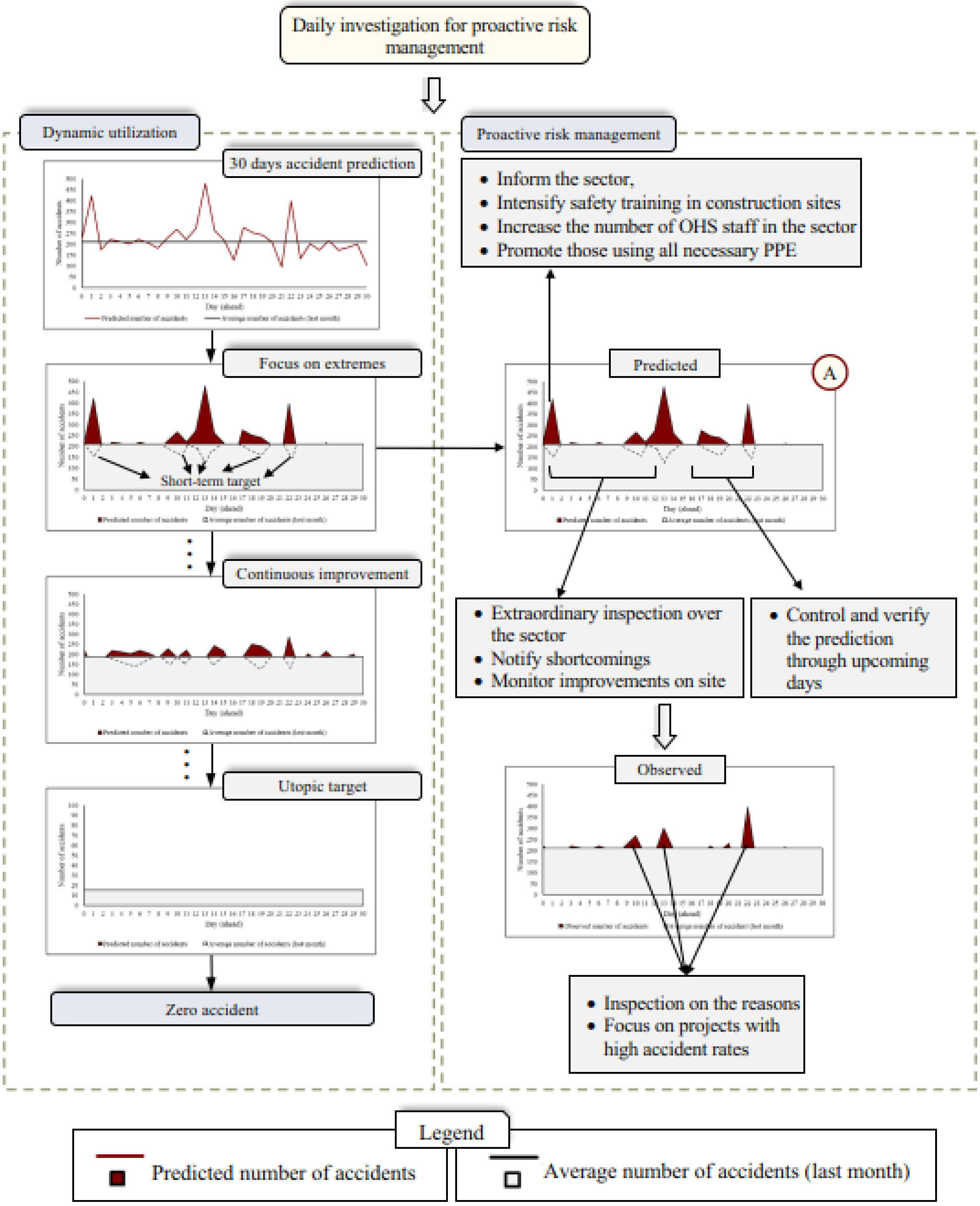

Daily investigation for proactive risk management
Dynamic utilization
30 days accident prediction
Focus on extremes
Short-term target
Continuous improvement
Utopic target
Zero accident
Proactive risk management
Inform the sector,
Intensify safety training in construction sites
Increase the number of OHS staff in the sector
Promote those using all necessary PPE
Predicted
A
Extraordinary inspection over the sector
Notify shortcomings
Monitor improvements on site
Control and verify the prediction through upcoming days
Observed
Inspection on the reasons
Focus on projects with high accident rates
Legend
Predicted number of accidents
Average number of accidents (last month)

Figure 14. Example of utilization process of the predictive model (Koc et al., 2022).

The authors described the model's utilization process in the Figure 14. One should focus on timing the actions in the model's predictions of the higher quantity of the safety hazards. Hence, the proactive safety actions are there for targeted and correctly timed the frequency of the hazards should decrease. If the enterprise can repeat this process over time the safety hazard rate is going to decrease continuously.

Oyedele et al. (2021) in *Deep learning and Boosted trees for injuries prediction in power infrastructure projects* argues that conventional machine learning techniques are not optimal in the modelling causes of injuries.

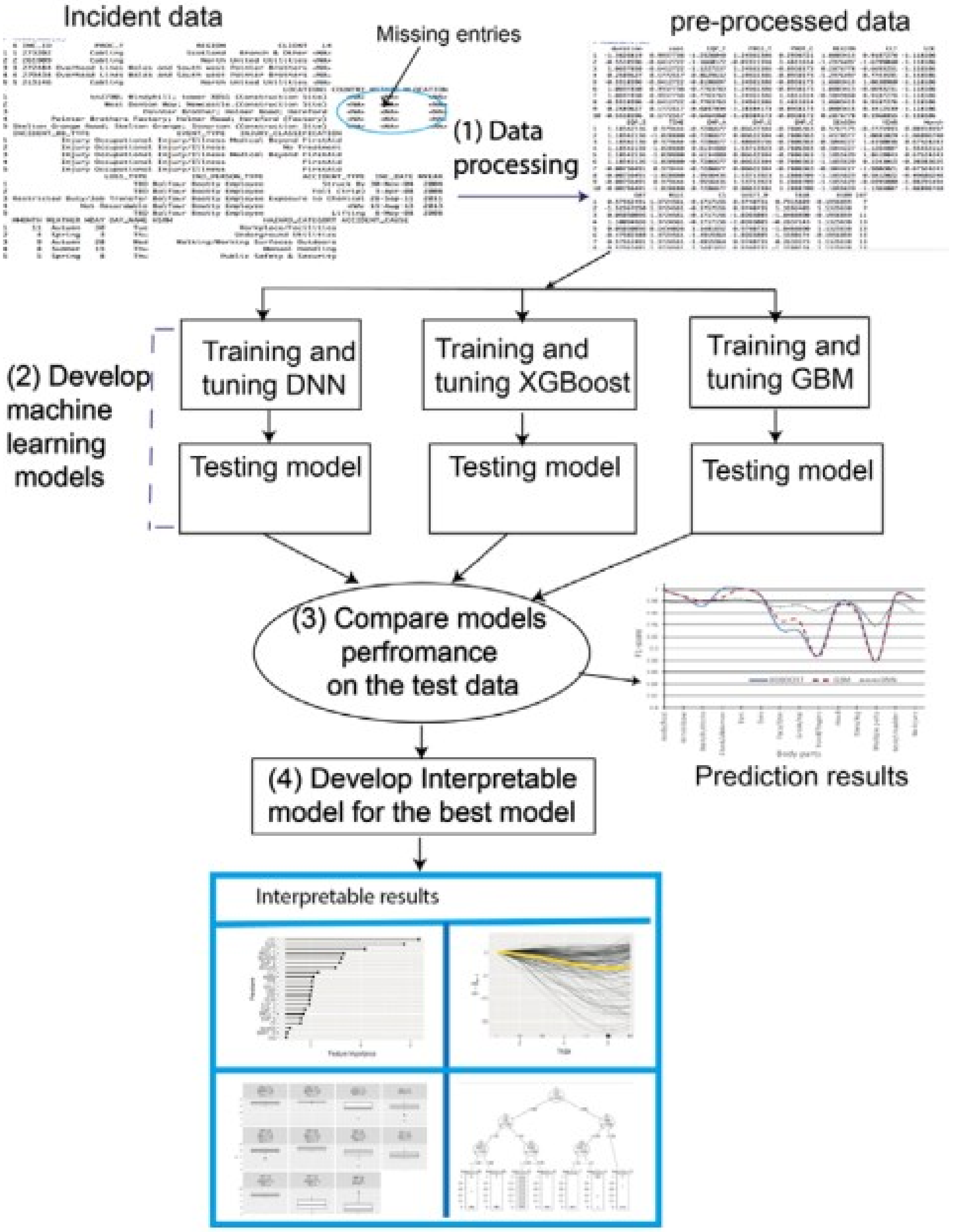


Figure 15. Process of the data preprocessing and ML model building for safety hazard prediction model (Oyedele et al., 2021).

The authors rely on deep learning and boosted trees, hence these models do not require manual data engineering compared to conventional models such as SVM and ANN. In the Figure 15 Oyedele et al. (2021) illustrates the model's process. The figure shows how data is preprocessed into structured form. After data handling machine learning models are developed and performance is measured with the test data set. After evaluation of the models the main model is tuned, and outputs are used for the safety hazard prediction. The authors recorded quite accurate prediction probability of 0,967 and Cohen Kappa rate of 0,964. The authors explained that according to the result of the sensitivity analysis deep neural network (DNN) based modelling has good generalization ability and the DNN has a good ability to identify connections between complex labels. The DNN neurons get quite high variation of the inputs that neurons handle through with an error rate, the error rate is used to modify the weight of the neuron in the model.

Oyedele et al. (2021) used DNN for the model. The predictors, such as site conditions or tool types, are used as predictors and the prediction of injured body part is done, see Figure 16. Also, the authors explained that with the model there is opportunity to predict likelihood for different combinations of the working domains. For example, the authors discussed how certain location and task indexes produced different probabilities for the head, ankle, and eye injuries. Some tasks and locations generate higher probability for the multiple injuries, the one could raise awareness and proactive actions for these tasks and conditions.

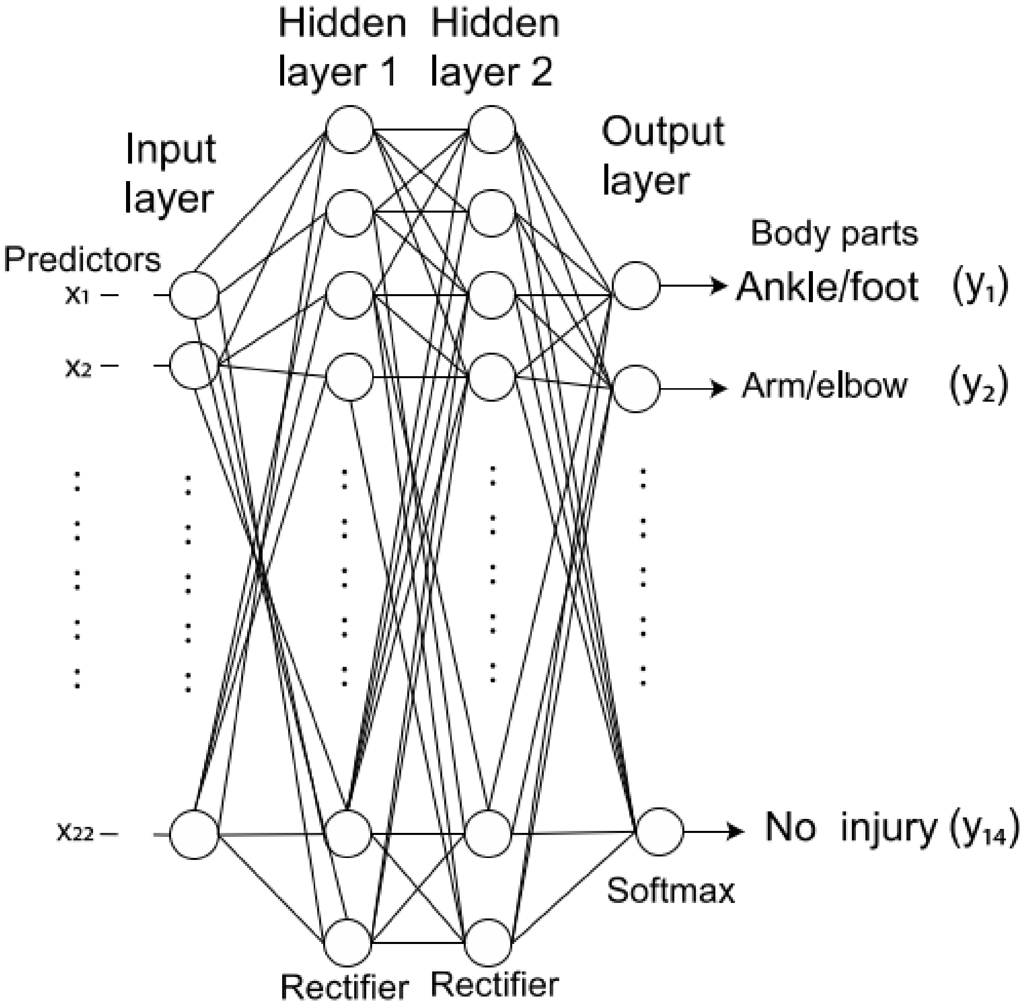


Figure 16. Illustration of the deep learning model's principle (Oyedele et al., 2021).

The authors explained that a task repetitiveness is linked to the injury rate of the linemen, a normal task seem to be more dangerous than rare task. The paper discusses the local interpretable model-agnostic that explains DL models predictions in the local scale such as hand injury. In practice, the model describes different predictors, labels, and effects of a prediction that the model produces. For example, the equipment used in the task has higher effectiveness in hand related incidents than the state of the electricity in the wire. The authors discusses that it might be that electrification causes more serious injuries while hand tool equipment rarely causes injuries to other body parts than hands.

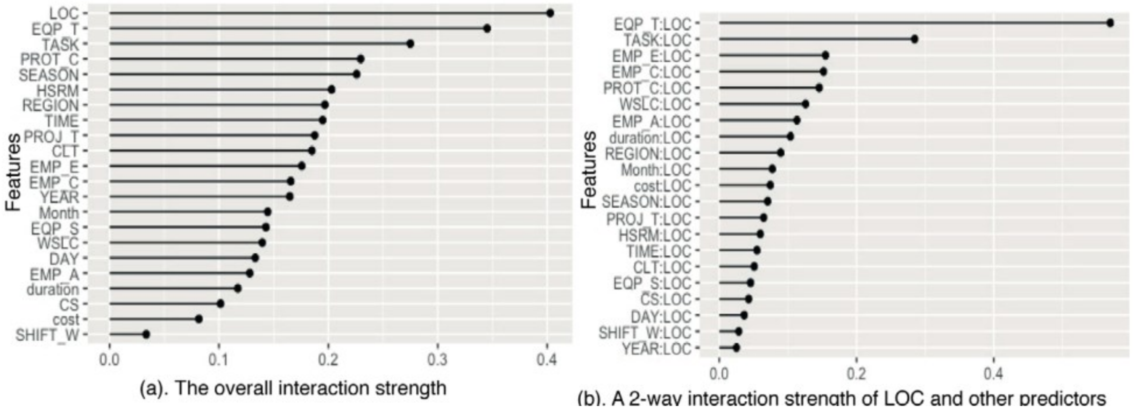


Figure 17. Interactions and relationships between the safety hazard features and predictions (Oydele et al. 2021).

Oyedele et al. (2021) explained that some project attributes affect to the incident rate more than others. For example, the site characteristics, equipment, and task type affect most in the incident occurrence, as can be viewed from Figure 17. This was confirmed by all models that the authors reviewed. The models recognized other powerful attributes, such as a season, project duration etc., from the data. The authors described the relationship between the predictors as an interaction strength of the predictor. The LOC predictor, that describes the site characteristics such as the terrain, ground conditions, wind conditions and site logistics and other characteristics of the site had the strongest interconnection to other attributes. The site conditions are, at the prediction domain, dynamic regarding to subject prediction of an incident. For example, the windy conditions are considered more dangerous to an eye than an ankle. This is simple logic and together with a machine learning model it perhaps has more importance than single observation.

In Deep Learning Models for Health and Safety Risk Prediction in Power Infrastructure Projects Ajayi et al. (2020) built six deep learning models together with the text-mining practices. In the Figure 18 the models are illustrated. In first stage the DNN model is producing the feed data for five other models in the stage two. In second stage the predictions are generated. The paper approached of the machine learning safety domain from a practical approach. The authors benchmarked their model with the existing models and developed user interface for better awareness of the safety issues. The models predicted relationship with different variables using mainly regression. The authors used area under curve, mean absolute error, kappa coefficient, sensitivity, and determination coefficient as the performance metrics. The authors examined different combinations of the layers and neurons compared to mean average error of their model, see Figure 19. This is logical approach for constructing optimal structure for the model. One should focus on structure where error's decline gradient is greatest.

The authors explain that the DNN's approach is divided into global and local views of the data. The global view is seeking an interaction between the predictors and the local view seeks explanation for the individual predictor's effect on the outcome.

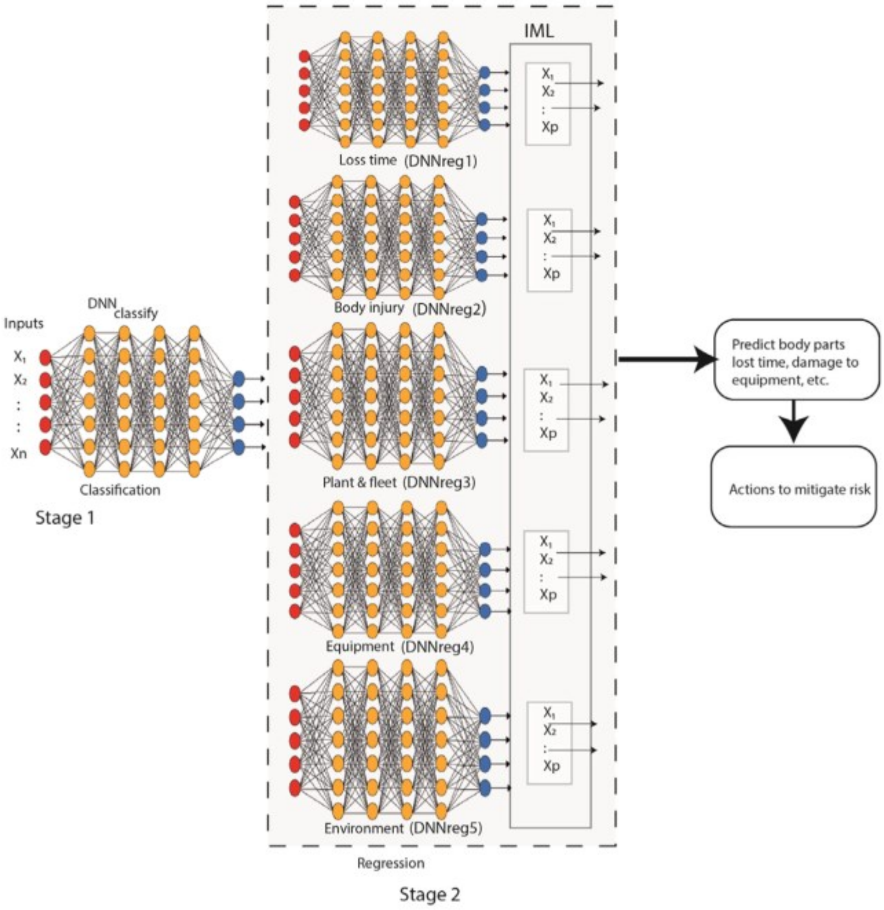


Figure 18. Illustration of the multi-stage DNN model (Ajayi et al., 2020).

The variables were determined with the text-mining approach from a health and safety incident cases of 17 972 that decreased to 16 900 at the set-up process. The variables extracted from the reports included, for example, condition of the personal protective equipment kit, project type and duration, and weather conditions.

According to the data that Ajayi et al. (2020) used, largest proportion of the injured body parts are fingers and backs with a cumulative proportion of 34% of all injuries. Hand, ankle, and knee injuries were cumulative proportion of 27%. It is noteworthy that 45% of the injuries were caused during excavation.

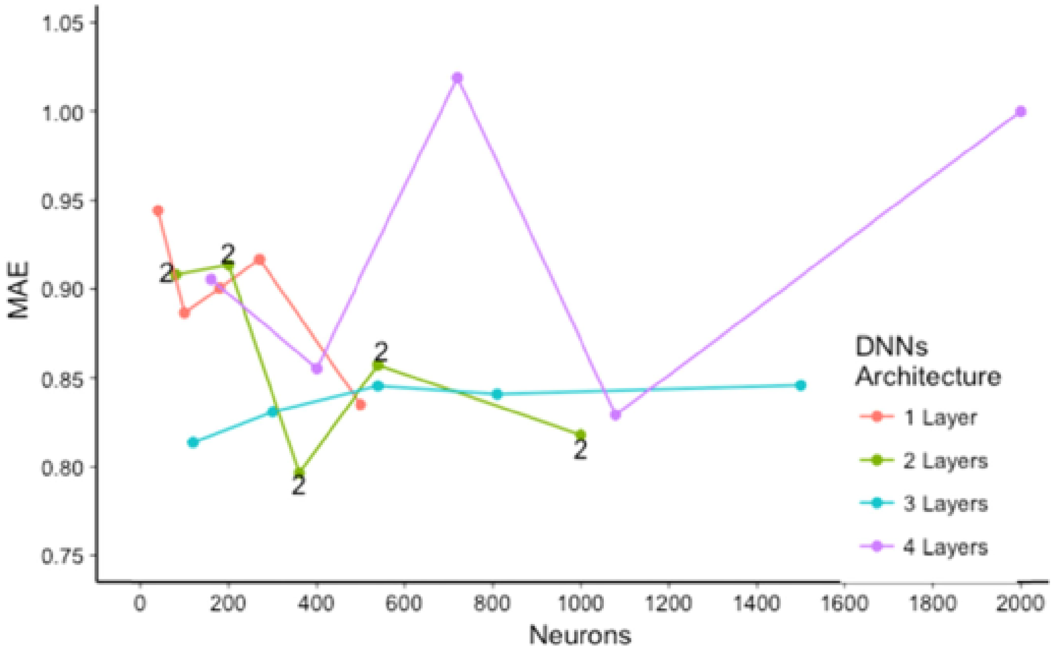
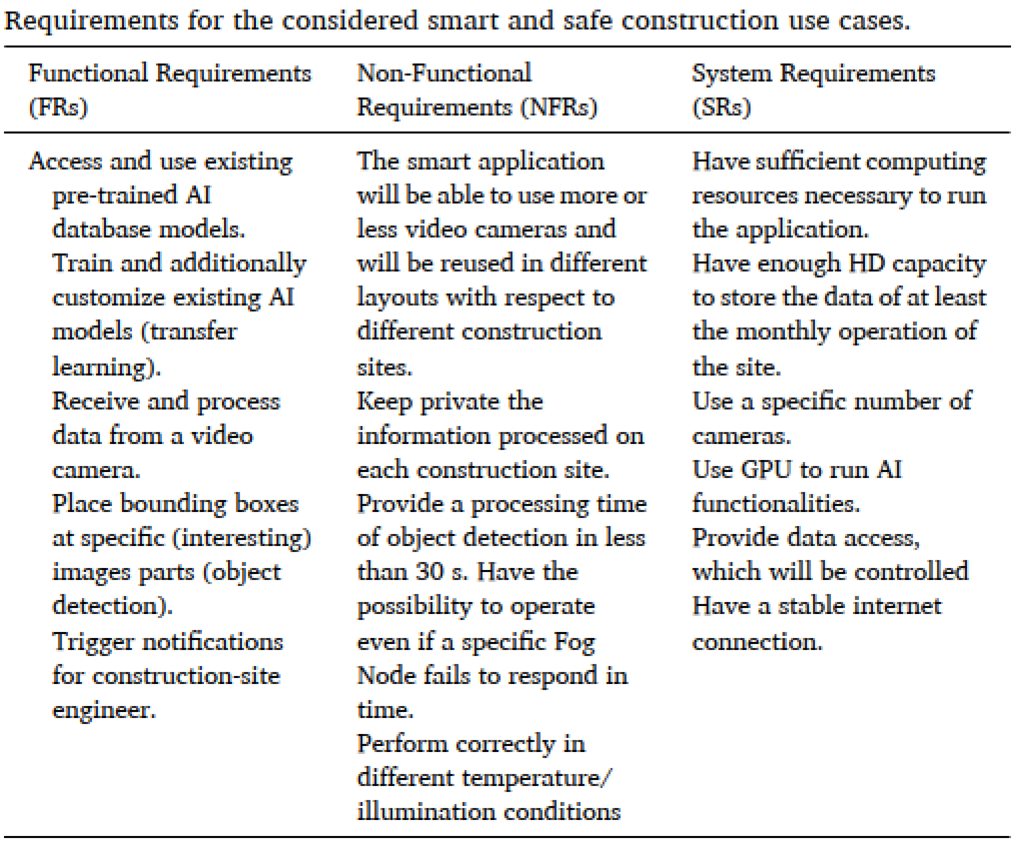


Figure 19. Relationship between MAE and number of the layers as function of neurons (Ajayi et al., 2020).

In *Building applications for smart and safe construction with the DECENTER Fog Computing and Brokerage Platform* Kochovcki and Stankovski (2021) discussed combining artificial intelligence, internet of things and blockchain technology. The authors constructed smart applications to produce the smart and safety construction sites. The authors focused to four different scenarios to achieve smarter and safer construction site. The scenarios included notifications for the site managers, surveillance of the site vehicles, management of the resources, assets and waste management and working conditions observation. The authors utilized a Decenter fog computing platform for the task, the platform is designed to run microservices needed in the smart and safety application utilization.



Requirements for the considered smart and safe construction use cases.

Functional Requirements (FRs)	Non-Functional Requirements (NFRs)	System Requirements (SRs)
Access and use existing pre-trained AI database models. Train and additionally customize existing AI models (transfer learning). Receive and process data from a video camera. Place bounding boxes at specific (interesting) images parts (object detection). Trigger notifications for construction-site engineer.	The smart application will be able to use more or less video cameras and will be reused in different layouts with respect to different construction sites. Keep private the information processed on each construction site. Provide a processing time of object detection in less than 30 s. Have the possibility to operate even if a specific Fog Node fails to respond in time. Perform correctly in different temperature/ illumination conditions	Have sufficient computing resources necessary to run the application. Have enough HD capacity to store the data of at least the monthly operation of the site. Use a specific number of cameras. Use GPU to run AI functionalities. Provide data access, which will be controlled Have a stable internet connection.

Figure 20. Different requirements for the smart and safety application layer for the construction site smart application (Kochovcki and Stankovski, 2021).

The Authors explained that end-users, such as construction engineers and managers gave positive feedback after the scenarios were utilized in the construction site. The authors note that a technical usability of the system supported information transformation from the site to application, and therefor for the users. In Figure 20 the authors go through technical requirements and principles of the smart application. The requirements are separated into functional and non-functional requirements together with the system level requirements. Functional requirements are related to system ability to achieve desired outcomes. Non-functional requirements set boundaries how the outcomes should be achieved. The system requirements are related to technical abilities of the system. Also, the authors note that system's ability to access different AI methods is important feature. It enables system to note, for example, personal safety equipment wearing issues.

Sattari et al. (2022) examined an AI-based decision-making system that takes assets in the consideration. The authors focused on the process safety management together with the asset management. The asset management were divided into two groups: assets and human resources.

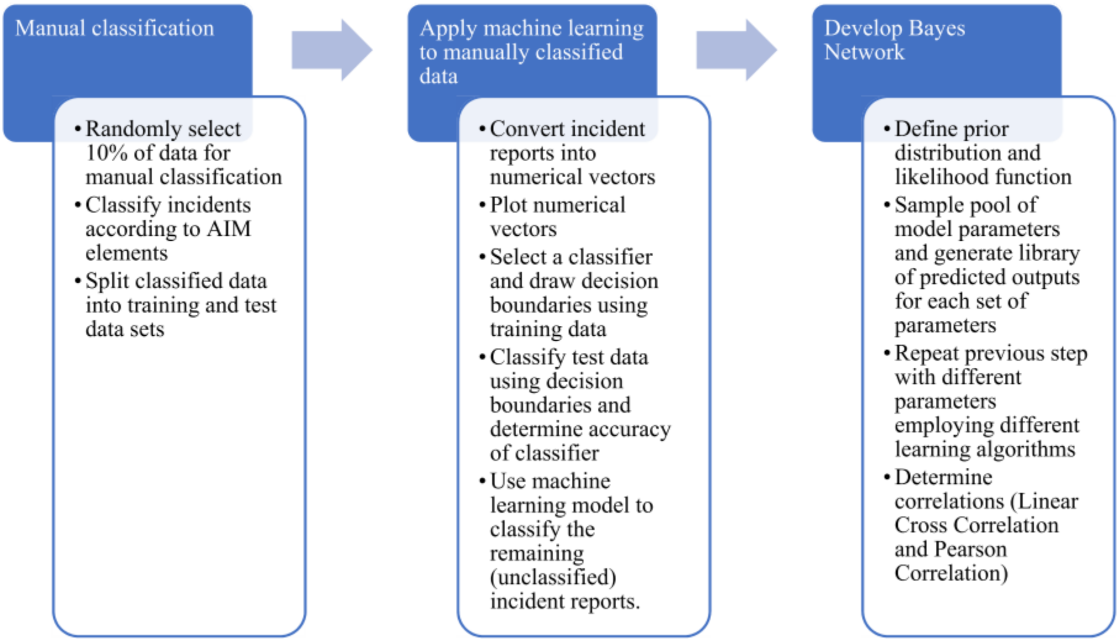

Manual classification
• Randomly select 10% of data for manual classification
• Classify incidents according to AIM elements
• Split classified data into training and test data sets
Apply machine learning to manually classified data
• Convert incident reports into numerical vectors
• Plot numerical vectors
• Select a classifier and draw decision boundaries using training data
• Classify test data using decision boundaries and determine accuracy of classifier
• Use machine learning model to classify the remaining (unclassified) incident reports.
Develop Bayes Network
• Define prior distribution and likelihood function
• Sample pool of model parameters and generate library of predicted outputs for each set of parameters
• Repeat previous step with different parameters employing different learning algorithms
• Determine correlations (Linear Cross Correlation and Pearson Correlation)

Figure 21. Example of the supervised machine learning process with process phase related tasks (Sattari et al., 2022).

In the Figure 21 the authors describe the process of the paper's machine learning modelling. In the manual classification phase part of the data are randomly selected for the classification that are done by the asset management process's elements. After classification the data is split for the machine learning part. The incident data is processed into numerical form and the classification model was built and data were classified for helping to develop the predictive model on the next phase. The authors selected random sample of 764 from the 7643 incidents. The authors describe that the classification problems are numerical problems, thus the plotted data can be divided with a decision boundary. Pythons TFidfVectorizer was used to generate vectorized data from the incident reports. The authors used linear support vector classifier to draw decision boundaries. The model uses the boundaries in the classification process. The authors searched for the dependencies and causes from the modified data. With the results of the paper the authors created a clear procedure and practical recommendations for each asset and operative class.

According to Pedregosa et al. (2011) the TFidVectorized is a Scikit-Learn based tool that transfers the terms occurring in the incident reports to numerical values. TFID is an abbreviation for times inverse document-frequency. The principle of the TDIF vectorization is to give more importance for the rarely occurring words because of the better information value of rare words.

4.2 AI utilization in project and portfolio management

Pan & Zhang (2021) discussed that a construction engineering and management benefit from the artificial intelligence in many ways. An AI can process multiple data sources that is beneficial for decision making. An AI can recognize patterns with a machine learning modelling and is able to do so with enormous dataset sizes. An AI is powerful tool for project and portfolio management. As discussed in the safety section the use domain of AI system is quite diverse and a user can achieve relatively high prediction rates with the machine learning modelling.

In *Big Data in the construction industry: A review of present status, opportunities, and future trends* Bilal et al. (2016) discussed the big data utilization in the construction domain. The paper reviewed multiple views on the big data domain such as data mining, data warehousing, machine learning and big data analytics. The authors discussed different practices that can be used with the big data related applications. A Document classification and analysis is used for, for example, classification documents to correct class and with document analysis the documents content can be examined.

In *Deep learning and Boosted trees for injuries prediction in power infrastructure projects* Oyedele et al. (2021) used project features, such as employee experience, project duration and project season for the prediction purposes. Similarly, the features can be used for a project and portfolio success prediction. The enterprise can perform the data analysis from a historical data and derive knowledge for the future purposes. This kind of action in the management process can lead the enterprise to more mature management process that is based on the knowledge from the past projects.

Mirnezami et al. (2020) concentrated on a project cash flow management together with a critical chain management and multi-criteria decision-making process. The authors divided the data to intervals that represents optimistic and pessimistic scenarios to gain knowledge from the data to project managers. The data represented, for example, most uncertain time of the project. The enterprise can use the information to allocate resources for more uncertain time of the project lifecycle, this seem to be rational approach in the multi-project domain. Hence, the managers time is limited it is rational to use energy for the most uncertain phase of the project.

In *Prediction of risk delay in construction projects using a hybrid artificial intelligence model* Yaseen et al. (2020) produced random forest classification model with genetic algorithm. The model's goal was to predict project delay problems at the construction business. The model gained accuracy of 91,67%, kappa of 87% and classification error of 8,33%. The authors explain genetic algorithm generates random decisions that produces mutation to the decision pool. Hence, the process is repeated the decisions are getting better in each round.

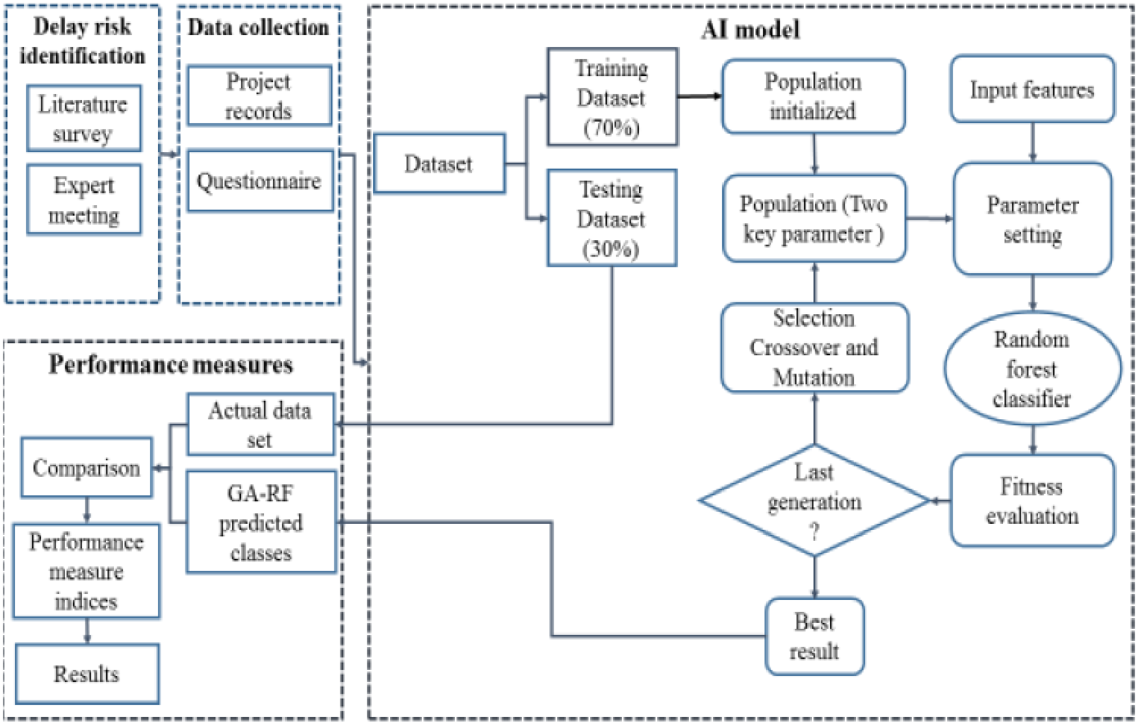


Figure 22. Delay risk identification process for predictive model building (Yaseen et al. (2020).

Yaseen et al. (2020) searched most common reasons for the schedule delays with the literature survey and expert meetings, see Figure 22. The data set were built based on the surveys and meetings. The model was feed with project features and RF classifier was set to predict projects schedule related risks. The genetic algorithm was used to tune parameters for achieve acceptable results. The reasons were categorized to seven groups, for example material and owner related purposes. The authors divided the reasons to the sub reasons and examined data of the 40 projects. The data was divided into risk levels together with a probability of occurrence. Also, the authors divided risk delay reasons to the classes by impact of the reason compared to original schedule and the model goal was to predict a delay category.

Chen and He (2012) did research on the cost management system that used data mining, the classification, cost analysis and cost forecasting. The authors explained that essential part of the data mining process is to form insights from the data through process of the machine learning. The data collection, data training, data testing and application derive knowledge for the future decisions. At the project classification the authors used decision tree model. The purpose was to identify suitable projects for the cost analysis phase. The cost analysis purpose was to clarify the cost structure of the projects. The authors explain that analysis is popularization and extension of principle component analysis. The analysis contained multiple horizontal and vertical layers, that covers an analysis of overall cost analysis together with geographic analysis.

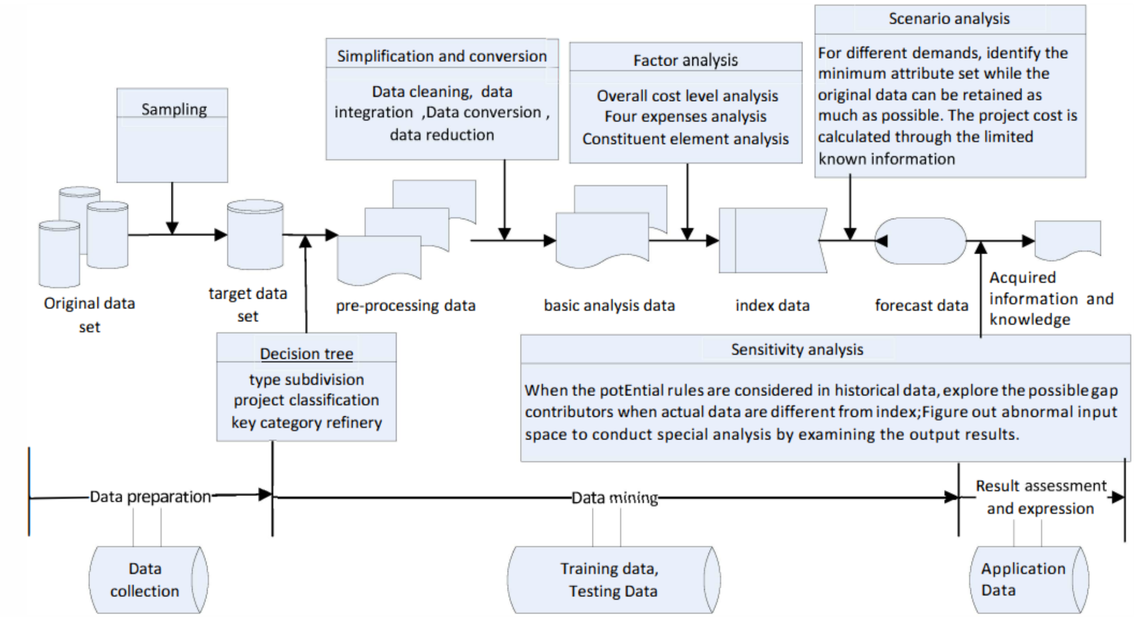

Sampling
Simplification and conversion
Data cleaning, data integration ,Data conversion , data reduction
Factor analysis
Overall cost level analysis
Four expenses analysis
Constituent element analysis
Scenario analysis
For different demands, identify the minimum attribute set while the original data can be retained as much as possible. The project cost is calculated through the limited known information
Original data set
target data set
pre-processing data
basic analysis data
index data
forecast data
Acquired information and knowledge
Decision tree
type subdivision
project classification
key category refinery
Sensitivity analysis
When the potEntial rules are considered in historical data, explore the possible gap contributors when actual data are different from index;Figure out abnormal input space to conduct special analysis by examining the output results.
Data preparation
Data mining
Result assessment and expression
Data collection
Training data, Testing Data
Application Data

Figure 23. The cost management prediction models' process activities (Chen and He, 2012).

Chen and He divided projects by project's building type, for example cabling projects and power distribution projects were separated, these categories had sub-categories for the more accurate analysis. In the Figure 23 the authors describe the actions for the data during the modelling process. After the data preprocess the authors executed factor and scenario analysis for the data. After forecasting the cost levels, the authors executed a sensitivity analysis for verifying the results before implementing derived knowledge further.

In *the Engineering Machine-Learning Automation Platform (EMAP): A Big-Data-Driven AI Tool for Contractors' Sustainable Management Solutions for Plant Projects* Choi, Lee, and Kim (2021) aimed to predict risks of the plant projects with a machine learning algorithm, also authors aimed to create a support decision system. The system consists of five modules; invitation to bid analysis, design cost estimation, design error checking, change order forecasting and equipment predictive maintenance see Figure 24. The information is gathered from existing plant's project data. The data is collected from enterprise resource planning systems together from commercial and public project data. Data were cleaned and preprocessed. In the machine learning platform, the data is further processed for machine learning basis. The authors used both regression and random forest together with the natural language processing activities. The authors note that existing solution, with also assisted by a machine learning model, demanded manual work from data analyst and experts.

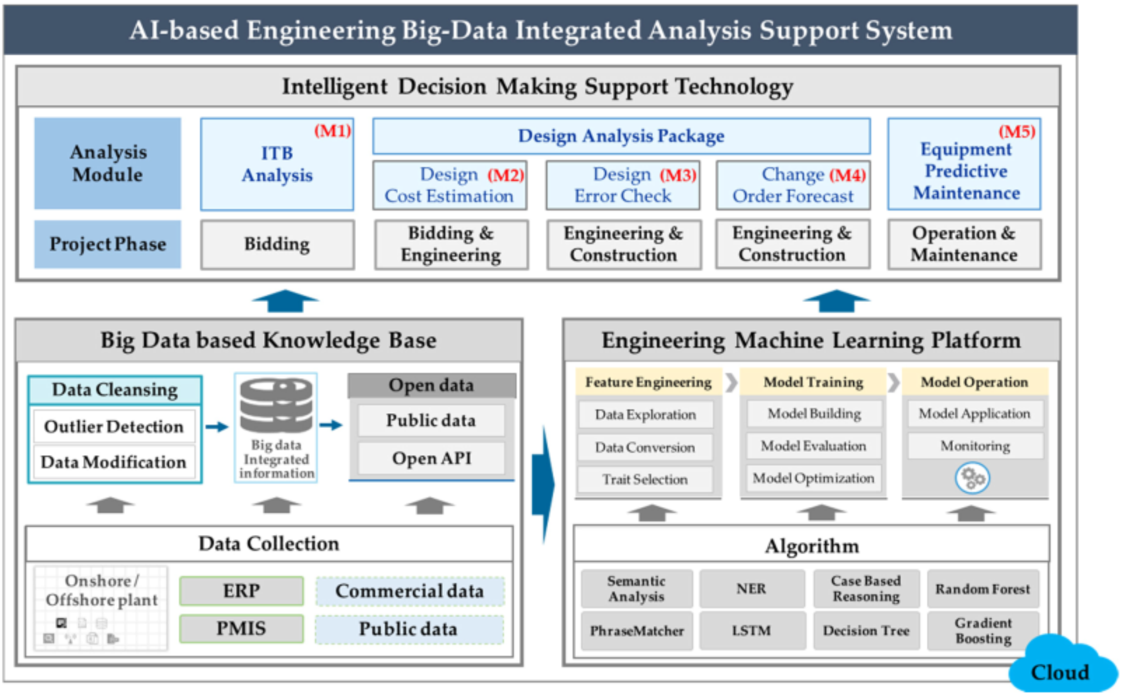
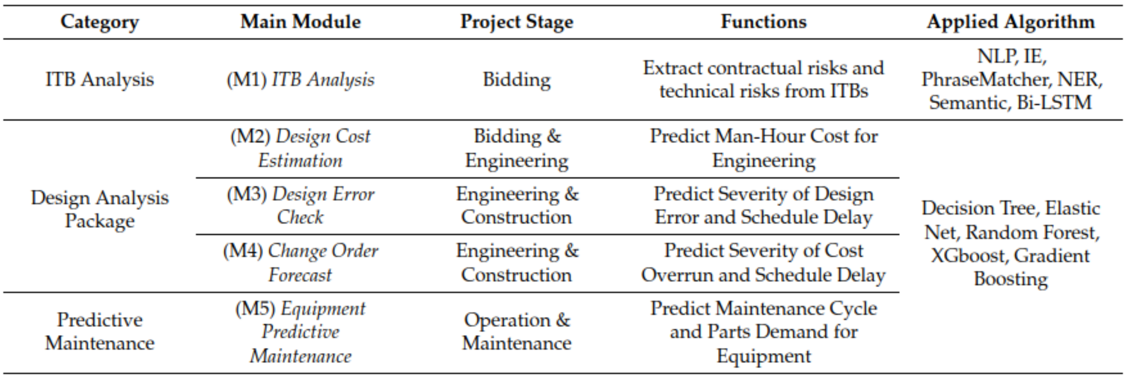


Figure 24. System architecture module illustration with data process of the artificial intelligence-based decision support tool (Choi et al., 2021).

The authors note that the data must be preprocessed. The authors used spaCy library for the text tokenization, lemmatization, POS tagging and dependency parsing. The model's process included a risk detecting part that used phrase matcher of the SpaCy library with fixed rules. The model can detect wanted keywords and phrases for preventing realization of the future risks. The authors built many submodules for different purposes such as direct clauses detection from a contract.



Category	Main Module	Project Stage	Functions	Applied Algorithm
ITB Analysis	(M1) *ITB Analysis*	Bidding	Extract contractual risks and technical risks from ITBs	NLP, IE, PhraseMatcher, NER, Semantic, Bi-LSTM
Design Analysis Package	(M2) *Design Cost Estimation*	Bidding & Engineering	Predict Man-Hour Cost for Engineering	Decision Tree, Elastic Net, Random Forest, XGboost, Gradient Boosting
	(M3) *Design Error Check*	Engineering & Construction	Predict Severity of Design Error and Schedule Delay	
	(M4) *Change Order Forecast*	Engineering & Construction	Predict Severity of Cost Overrun and Schedule Delay	
Predictive Maintenance	(M5) *Equipment Predictive Maintenance*	Operation & Maintenance	Predict Maintenance Cycle and Parts Demand for Equipment	

Figure 25. Modules and functions of the AI decision support application by (Choi et al., 2021).

The authors examined various EPC projects for defining suitable keywords to determine relationships to the risks. The risks were determined by the impact of the risk to strong, moderate, and weak impact.

In Figure 25 thew authors describe the purpose of the modules together with their functions. Also, authors note algorithms used for separate modules. For the plant project bidding phase, the algorithms are focused on natural language processing and other modules are mostly focused on the numerical processing. One can view this also from the Figure 26, the paper’s process is divided into two segments. The authors note that the model’s ability to achieve simultaneously high precision and high recall results is limited. For example, highest contrast at the model’s prediction were recall of 99,3% and precision of 54%. The authors explain that low precision in the risk detection submodule is explained by duplicate risk detection.

Choi et al. (2021) produced decision making system using the data from the selected projects. Also, the authors utilized the machine learning for the different submodules of the system. The system collects risks from the documents, such as bidding documents. The system predicts design costs together with quality of the design in terms of errors and schedule. Also, the system is set to predictively maintain parts and equipment. The authors note that the system was produced by using python. The performance of the system's prediction rates by F1 measure were 70%, 86,8%, 87,6%, and 88,4%. The authors note that the system integration into cloud-based system benefits the enterprise by providing discussed applications easily on site. The system can detect risk through the project's life cycle; thus, the risk level is relatively lower during the project execution. The author's intention was to produce a system that requires low or no machine learning experience during usage. Also, the automated risk analysis requires fewer work hours, thus the project management can use more work hours to manage the projects. Hence, the process is automated risk management is done to every project at the same level. This level could be considered at the minimum level of the risk management.

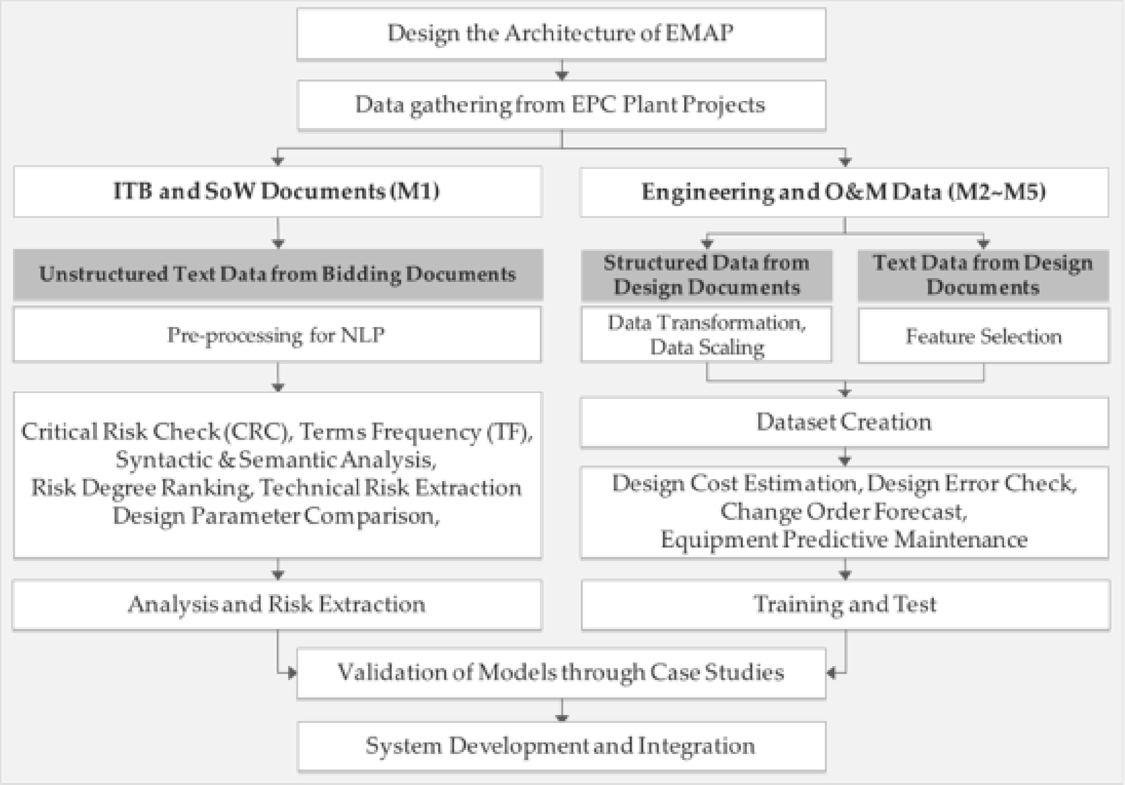

Design the Architecture of EMAP
Data gathering from EPC Plant Projects
ITB and SoW Documents (M1)
Engineering and O&M Data (M2~M5)
Unstructured Text Data from Bidding Documents
Structured Data from Design Documents
Text Data from Design Documents
Pre-processing for NLP
Data Transformation, Data Scaling
Feature Selection
Critical Risk Check (CRC), Terms Frequency (TF), Syntactic & Semantic Analysis, Risk Degree Ranking, Technical Risk Extraction Design Parameter Comparison,
Dataset Creation
Design Cost Estimation, Design Error Check, Change Order Forecast, Equipment Predictive Maintenance
Analysis and Risk Extraction
Training and Test
Validation of Models through Case Studies
System Development and Integration

Figure 26. The model's development process (Choi et al., 2021).

5 Process for AI utilization

Building AI-based systems demands resources from the enterprise. In *Challenges of data refining process during the artificial intelligence development projects in the architecture, engineering and construction industry* Heo et al. (2021) discussed human resources that an AI development project need, the authors examined need of the human resources from the qualitive, and quantitative perspective and tasks related to the project.

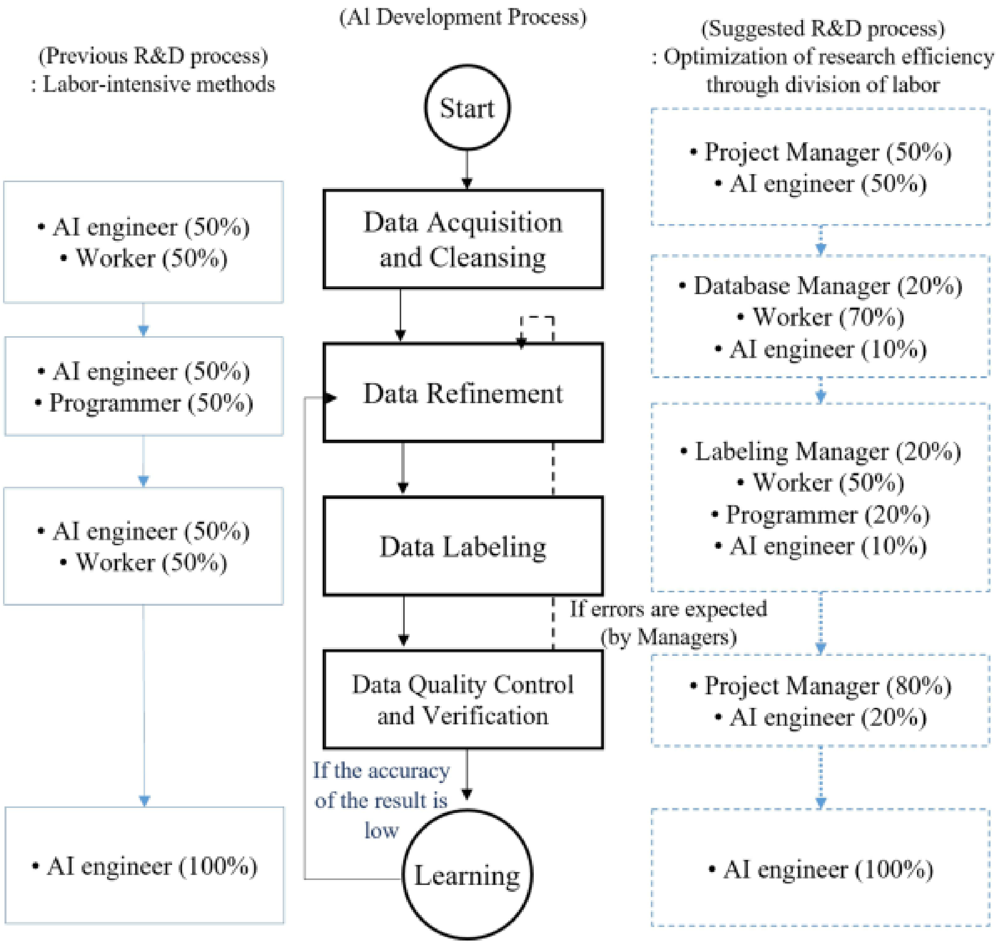


Figure 27. The AI process and tasks divided into work positions (Heo et al., 2021).

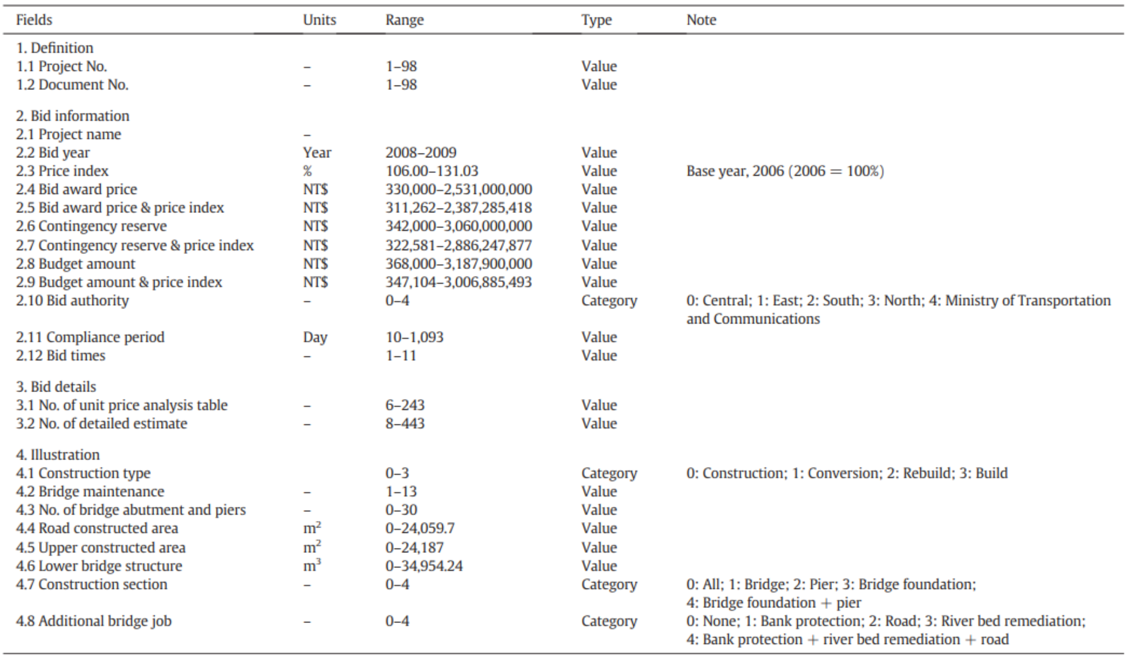
Heo et al. (2021) notes that developmental projects with an AI should has continuous data modification process. The first modification is a start of refinement process where the data evolves to better form. The authors discuss that the data can be divided into two categories, image-based and time series data. The authors describe that process of data utilization demands collection of the raw data, data modification, segmentation and labeling. After data modification a machine learning model is introduced with the data and tuning the model begins. During the process the quality control is done for achieving optimal results. This process is also illustrated in Figure 27 together with the workflow related work positions and proportions of work times. The paper concluded that project management's participation from the beginning of the project is beneficial for the later phases.

Heo et al. (2021) represented a work index for the evaluation of needed resources for the AI project, see equation (6). It is notable that one should concentrate among the data amount to the data quality. Complexity of the raw data and demanded results for the data refinement and machine learning models' performance depends on the data quality.

$$Work\ index = \frac{Total\ amount\ of\ a\ data}{Degree\ of\ input\ manpower \cdot Work\ hours} \tag{6}$$

The authors executed a case study where they observed results of the model when the refinement manager was used to manage the research and development team. Thus, the AI expert could focus on the model and data refinement. Also, the manager can execute the quality control on the process and data.

In *Optimized artificial intelligence models for predicting project award price* Chou et al. (2015) strived to forecast a project price with AI modelling. The price of the project evolves quite rapidly, hence the changes in costs of construction. The authors used multiple approaches; multiple regression analysis, artificial neural network, and case-based reasoning. The authors examined bid materials of the near 100 bridge projects. The details of the used data in Figure 28.



Fields	Units	Range	Type	Note
1. Definition				
1.1 Project No.	–	1–98	Value	
1.2 Document No.	–	1–98	Value	
2. Bid information				
2.1 Project name	–			
2.2 Bid year	Year	2008–2009	Value	
2.3 Price index	%	106.00–131.03	Value	Base year, 2006 (2006 = 100%)
2.4 Bid award price	NT$	330,000–2,531,000,000	Value	
2.5 Bid award price & price index	NT$	311,262–2,387,285,418	Value	
2.6 Contingency reserve	NT$	342,000–3,060,000,000	Value	
2.7 Contingency reserve & price index	NT$	322,581–2,886,247,877	Value	
2.8 Budget amount	NT$	368,000–3,187,900,000	Value	
2.9 Budget amount & price index	NT$	347,104–3,006,885,493	Value	
2.10 Bid authority	–	0–4	Category	0: Central; 1: East; 2: South; 3: North; 4: Ministry of Transportation and Communications
2.11 Compliance period	Day	10–1,093	Value	
2.12 Bid times	–	1–11	Value	
3. Bid details				
3.1 No. of unit price analysis table	–	6–243	Value	
3.2 No. of detailed estimate	–	8–443	Value	
4. Illustration				
4.1 Construction type		0–3	Category	0: Construction; 1: Conversion; 2: Rebuild; 3: Build
4.2 Bridge maintenance	–	1–13	Value	
4.3 No. of bridge abutment and piers	–	0–30	Value	
4.4 Road constructed area	m^2	0–24,059.7	Value	
4.5 Upper constructed area	m^2	0–24,187	Value	
4.6 Lower bridge structure	m^3	0–34,954.24	Value	
4.7 Construction section	–	0–4	Category	0: All; 1: Bridge; 2: Pier; 3: Bridge foundation; 4: Bridge foundation + pier
4.8 Additional bridge job	–	0–4	Category	0: None; 1: Bank protection; 2: Road; 3: River bed remediation; 4: Bank protection + river bed remediation + road

Figure 28. Informative attributes of the project bidding data (Chou et al., 2015).

The authors evaluated the model's performance with mean absolute percentage error. The genetic algorithm together with the artificial neural network model achieved average MAPE rate of 7,526%. The genetic algorithm is a search algorithm that uses stochastic approach. The algorithm sets weights in the model's calculation and weights are tuned during the process.

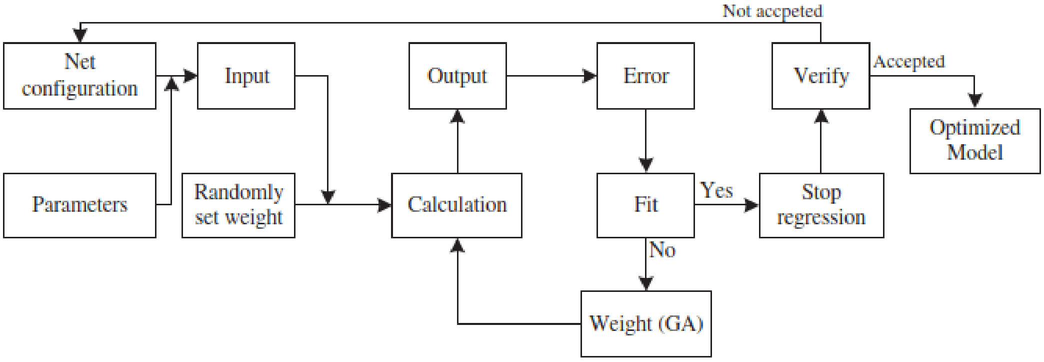


Figure 29. Model's prediction development information flow (Chou et al., 2015).

The authors conclude that data points with numerical values tend to have higher correlation to the project award amount than categorical values. Although the authors explain that categorical values can introduce regional differences between project bids. In Figure 29 process of the model's development. In the beginning configuration is set together with the parameters. The parameters affect to project prices and are tuned by the genetic algorithm. After acceptable level of error has been reached the model's performance is verified and further optimized.

6 Conclusion

In sections 2 and 3 the review on the present applications was made. In section 2 the review was made for the safety domain. Section 3 consist of the review on the project and portfolio management domain.

In the section 2 different approaches for utilizing the AI to safety related purposes has been reviewed. Practically all papers aimed to solve safety incident causes for raising knowledge to future development purposes. Some authors used the information for developing the end-product, such as a risk warning system or smart application. According to the review it is common to use multiple predictive models. Considering the work done in the data preprocessing phase it is logical to formulate a new model continuing of the previous model. Natural language processing together with classification- and regression-based models were used as the tools for achieving ability to predict incident causes and other factors that affected in occurrence of the incidents.

Baker et al. (2020) focused on the key elements of the site condition during the occurrence. Similarly, Oyedele et al. (2021) examined relationships between the incidents and site conditions. They also investigated the relationship between the incidents and work phase and tools used during the incidents. Zhang et al. (2019) classified the causes of the incidents with five model combination that was tuned by the SQP algorithm. Liang and Liu (2021) were keen to solve three main elements affecting to the occurrence of the incident, the authors also aimed to formulate repeatable conditions for the system. This aimed to effect on the quality of the predictive model. Thus, the input conditions are kept similar the model performs better. Goh and Ubeynarayana (2017) exploited preset labels during the incident cause prediction process. Pan and Zhang (2021) focused on identification of the risks related to different work phases related to the incidents. Identification of the relationship helps project personnel to time predictive actions correctly. Cheng et al. (2020) used natural language processing activities to unravel causes of the incidents from the accident report narratives. Ajayi et al (2020) focused to solve the relationship between the incident reason and the injured body part. The only machine vision-based paper reviewed was by Arashpour et al.2022). The paper's purpose was to demonstrate how machine can define position of an excavation machine. Kochovcki and Stankovski (2021) were keen to build smart application for construction site's use, the application purpose in safety domain was to generate notifications for site personnel. For example, detecting PPE use through the site video monitoring. The application of the risk warning system was examined by Lin et al. (2021). Koc et al. (2022) viewed occurrence of the incidents with the time series examination, the paper focused on predictive actions with knowledge of predicted incidents based on the time series. Sattari et al. (2022) focused on safety in the process and asset management domain. The authors main emphasis was to find relationships between different incidents and different assets and processes.

The review focused on the project management domain in the section 3. Bilal et al. (2016) reviewed mainly classification related model usage in construction industry, such as classification of a document into right class. Mirnezami et al. (2020) examined ways to generate scenarios of the project outcome with project related data, the authors discussed about project cashflow, critical chain management and supportive multi-criteria decision-making system. The attributes that Oyedele et al. (2021) used for the safety incident relationship prediction can be used for the project managemental purposes. Also, it is reasonable to seek the attributes that effect directly on the project outcome from project managemental view. Yaseen et al. (2020) examined most common reasons for a project delay which were gathered from the expert meetings. The reasons were used to model the projects schedule risk levels and occurrence probabilities. Chen and He (2012) used decision tree for finding suitable projects into the cost analysis phase. The authors also noted that it is valuable to explore the data hence this can give valuable insights itself, the modelling demands one to process the data. Choi et al. (2021) used enterprises resource planning system's data together with the existing project data for reducing the risks on the project, the authors produced system with multiple modules that focused on the separate phases, such as a design review, of the project.

6.1 Data process and model

The data should be organized with the way that supports later phases of the data exploration. Also, one can collect data from the multiple sources and merge the data. This could lead for increased knowledge of attributes that business has generated over time. One should explore through the data and process the attributes. For example, one can pivot the data table to the wanted format, handle the missing values, and calculate means and averages for reducing the variance of the attributes. One can aggregate lower-level information data to upper-level information, such as an information of the municipality to region level. Also, one can normalize data. The normalization reduce effects of extreme values from the data by normalizing the values for determined range.

A complexity and structure of the data determine the procedure for the data. Also, a problem formulation affects to demanded procedure. The machine learning domain is strongly connected to the statistical problem solving and Python with machine learning libraries is often used tool. The Python's popularity has driven creation of the useful resources around the platform that helps user to solve individual problems with a low amount of the experience. It is noteworthy that although the data processing and the model building is time consuming despite the relatively simpleness of the coding with the Python or similar coding environment.

The data processing is problem dependent and highly related to the data's structure that is processed. A precise representation of the data processing activities is therefore quite inconvenient task to execute. Nevertheless, data preprocessing is an important task to execute properly. One can handle missing values, for example, by filling the values with an average of missing attribute value or by deleting a data related to missing attribute. Also, one can process, encode, the data from verbal form to numeral form. The machine learning algorithms understands latter better. For example, one can encode the categorical string type data from the safety report to numerical form that allows the algorithms to process the data. One can also handle data observations that differs significantly from other observations by detecting and cleaning these observations. This is called outlier detection. The outliers are problematic for statistical calculations, such as mean calculation. The outliers distort the calculations.

As one can view from Figure 27 the AI's learning process has more weight on data processing phases compared to actual learning phase. One can view from the figure that 80% of estimated working hours are used before learning phase. My experience supports this view hence I would estimate that 90% of the time is used within data processing. As mentioned earlier this is quite good opportunity to generate insights of the business domain by going through data related on the business operations.

6.2 Safety management – Data utilization for safety development

According to the review there is various ways to utilize the AI in the construction business domain. A logical approach is to start utilizing the data that already exists in the enterprise's database. After the utilization one can have better understanding on the data's abilities, strengths, and weaknesses. After this the enterprise can modify the data collection process towards demanded data attributes and formalities.

After general data cleaning and preprocessing the text mining activities could be executed on the safety incident data's descriptive text part. For the other attributes, such as the day of the week, time, project stage and other numerical or categorical values the visualization could be done. One can seek for the relationship between operative attributes and project life-cycle related attributes. This could gain knowledge for the proactive actions. The knowledge could also develop the risk management in domains safety and project management.

The relationships between project stage, project season, tools and incident time were examined in the safety section. Also, one can seek for asset related relationships between the incidents and other attributes. In the data preprocessing and visualization phase one can formulate insights from the data, the insights can be highly valuable for the enterprise.

After the data is processed, examined, and visualized one shall build model for generate, for example, predictions from the data. Also, one can use preprocessed data for the diagnostic use. Diagnostic approach can explain the data and determine proactive actions to prevent similar incidents. Choosing the model is also dependent on the problem and dataset's attributes that is under examination, and it is quite difficult to outline a best model for all purposes. According to the review, the use of the multiple models is a common approach. One can also consider using the outperforming optimized model as the chosen one. Hence, the data processing is the most time-consuming phase, it is logical to rely on the multiple models. Hence, multiple models can be tested one can select the best model in terms of accuracy for the selected objective. Furthermore, one can optimize the model for achieving better results.

6.3 Project management – risk and activity management

The main goal of the project and project management personnel, such as a project manager, is to achieve planned cost, schedule, and scope. The goals are related to, for example, different processes, interactions between stakeholders, resource management, and leadership.

The data preprocessing is valuable for different operative domains. Classification of the documents, and content of them, could be beneficial for operative use of the build networks. Also, a quality of the documents is beneficial in the later project management manner. Thus, the projects often connect each other from a scope's perspective but not from a schedule perspective. If the content of the documents is structured, or semi-structured, the machine should be able to determine the quality of the content. It is notable that this works for safety document domain also. One can determine the relationship between the incidents and the safety documents related to the work.

The design and maintenance of the cash flow is often important part of the project management. The cash flow often follows the milestones of the project. Thus, the milestones usually connect most important points of the projects to practical work or deliveries it is logical to manage the cashflow in regular basis. If the cash flow is managed regularly a project cost, scope and schedule are more predictable.

One can also determine the most common reasons for different deviations of the projects. When a machine learning model can detect deviations the risk occurrence is possible to solve. Project activities, such as design, can be measured from the design output but also from the numerical attributes. Thus, in the distribution networks the design's output is affecting to the scope and therefor the budget of the project. Measurement for every project lifecycle points, such as bidding, starting the project, during the design could lead to quality in terms of managing deviations.

6.4 Summary

The next practical steps are related to data availability in the present time. The data could be used for generating new insights during the data processing and from, for example, the predictive model's outputs. The focus should be on the safety and project management domain because these are the key elements in the Elenia's construction domain. As explained in sections 6.2 and 6.3 Elenia should utilize existing data for artificial intelligence use. The enterprise operating domain and management requires predictions that are used for the project and safety management. Also, the predictions are used for decision making purposes and are therefore essential for the enterprise's operation. Also, considering the literature review's outcome and the domain where Elenia is operating the machine learning approach is a logical choice.

The literature review on AI utilization in construction domain optimistically drives future development in the enterprise. The review gave practical use cases for the further development purposes. Hence, it demands less experiment, and therefore, less resource when one elaborates existing solution compared to producing a new solution. Thus, it is logical to approach AI utilization with the literature review's examples. As the review's outcome, there is no silver bullet for producing machine learning systems that work in every operational domain. Thus, the model's ability to achieve an objective is dependent on the problem and the data structure. The cases should be considered one by one. Furthermore, the review's use cases exemplify how AI and machine learning domain can be used in Elenia's construction domain.

It is noteworthy that even the machine learning models may seem to be superior on the tasks they are built for, the generalization between unrelated task is unknown. In other words, the ability to solve different problems than the model is made for may not work. Also, the data structure and attributes effect the possibilities of how one can utilize the data and AI. One may have to use significant resources for the preprocessing of the data. Also, most of the review's use cases focused on supporting a decision-making process with a machine-learning solution. Therefore, it is logical to develop systems for supportive decision making. As mentioned, the dependency on the problem and data demands resources from the enterprise. Thus, it is important to identify the tasks where machine learning potential is greatest. According to the review, the probability of achieving the benefits from AI utilization is greater when the data's size is significantly large and there is a reasonable number of attributes that correlate with the outcome. The machine learning solutions may not work properly when the data's structure or size is inconvenient. Also, the solutions are often built for a specific need. Therefore, the solutions are not performing well if the outcome is not clearly described.

It is quite easy to understand the opportunities of the utilization of the AI and machine learning, for example, to safety management domain. Hence, a machine learning model can describe what elements are related to different outcomes one can use this for deciding what proactive actions could be made. For example, the machine learning model can collect simultaneously information from the enterprise’s resource management program, Finnish meteorological institute’s data, safety incident data. Also, the model could simultaneously describe, for example, the safety material usage and updates and formulate a risk level for incident by using all that information. A project personnel, such as project manager and safety supervisor, can risk level detection to prioritize time and proactive actions to projects with higher risk level. A strength of the machine learning models is ability to collect and process enormous amounts of the data. Also, the analysis process is continual and similar. Thus, the model’s actions are high in quality.

The next steps for utilization Elenia should take advantage of the existing data. The enterprise can process the data for formulate deep insights from key processes. After the data is examined, one can define the key elements and actions for achieving the objectives. After the process one can define, with the assistance of machine learning, the relationships of the key elements to outcome of the project. It is noteworthy that the outcome contains the safety measurements of projects. The enterprise should develop process parts that effect into key processes. Over time, the development of key elements could be used for building decision support system that has multiple modules, such as project plan and safety modules. The system could have multiple approaches to take advantage of AI. Previously built models for the safety and project management domain could work as the modules in the system. The machine vision- and sensor-based warning systems are also interesting from the enterprises operating domain. For example, one could increase the safety of worksite if a machine vision could be utilized for warning undesirable movement by excavator near high voltage overhead lines. Hence, these systems demand equipment that may be expensive to obtain in the work sites. Also, equipment must be connected to artificial intelligence-based system with real-time connection in demanding conditions. The requirements raise technical demands for the system. Hence, most of the challenges related to safety and project management can be solved by integrating available information efficiently in the processes. It is logical to start development from the knowledge and data utilization. The applications can support enterprises’ personnel to follow the determined processes and instructions punctually.

The Thesis’s focus was on the literature review on the selected domains for utilizing the AI. The safety, project and portfolio management domain were examined for collect information related to the opportunities that AI offers for Elenia’s operational domain. Optimistically, Thesis’ acts as one of the first cornerstones for significant utilization of AI in Elenia’s construction business. Hence, Thesis’ included only the literature review, the practical approaches are apart from the Thesis. For future works, the technical implementation of machine learning models could be examined. As mentioned, this Thesis´ focused on the utilization of AI with machine learning applications. Thus, the review on the other applications of AI utilization could be beneficial for the enterprise’s process development.

References

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of machine learning research*. https://doi.org/10.5555/1953048.2078195

Baker, H., Hallowell, M. R., & Tixier, A. J. (2020). AI-based prediction of independent construction safety outcomes from universal attributes. Automation in construction, 118, 103146. https://doi.org/10.1016/j.autcon.2020.103146

Storopoli J., Huijzer R. & Alonso L. (2021). Julia Data Science. https://juliadatascience.io. ISBN: 9798489859165.

Bezanson J., Karpinski S., Shah V. B. & Edelman. (2012). Why We Created Julia. Retrieved 20, September. https://julialang.org/blog/2012/02/why-we-created-julia/

Romano, F. (2018). *Learn Python programming: A beginner's guide to learning the fundamentals of Python language to write efficient, high-quality code*. Packt Publishing.

Skorpil, V., Oujezsky, V., Cika, P. & Tuleja, M. (2019). *Parallel Processing of Genetic Algorithms in Python Language*. https://doi.org/10.1109/PIERS-Spring46901.2019.9017332

Mezher, M. A. (2022). PGFLibPy: An Open-Source Parallel Python Toolbox for Genetic Folding Algorithm. *Journal of advanced computational intelligence and intelligent informatics, 26*(2), 169-177. https://doi.org/10.20965/jaciii.2022.p0169

Zhang, F., Fleyeh, H., Wang, X. & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in construction, 99*, 238-248. https://doi.org/10.1016/j.autcon.2018.12.016

Goh, Y. M. & Ubeynarayana, C. (2017). Construction accident narrative classification: An evaluation of text mining techniques. *Accident analysis and prevention, 108*, 122-130. https://doi.org/10.1016/j.aap.2017.08.026

Pan, Y. & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Automation in construction, 122*, . https://doi.org/10.1016/j.autcon.2020.103517

Zhou, Z., Goh, Y. M. & Li, Q. (2015). Overview and analysis of safety management studies in the construction industry. *Safety science, 72*, 337-350. https://doi.org/10.1016/j.ssci.2014.10.006

Rauhiainen, L. & Estra, C. (2018). Artificial intelligence: 101 things you must know today about our future. [Lasse Rouhiainen].

Cheng, M., Kusoemo, D. & Gosno, R. A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in construction, 118*, 103265. https://doi.org/10.1016/j.autcon.2020.103265

Arashpour, M., Kamat, V., Heidarpour, A., Hosseini, M. R. & Gill, P. (2022). Computer vision for anatomical analysis of equipment in civil infrastructure projects: Theorizing the development of regression-based deep neural networks. *Automation in construction, 137*, . https://doi.org/10.1016/j.autcon.2022.104193

Koc, K., Ekmekcioğlu, Ö. & Gurgun, A. P. (2022). Accident prediction in construction using hybrid wavelet-machine learning. *Automation in construction, 133*, 103987. https://doi.org/10.1016/j.autcon.2021.103987

Oyedele, A., Ajayi, A., Oyedele, L. O., Delgado, J. M. D., Akanbi, L., Akinade, O., . . . Bilal, M. (2021). Deep learning and Boosted trees for injuries prediction in power infrastructure projects. *Applied soft computing, 110*, 107587. https://doi.org/10.1016/j.asoc.2021.107587

Ajayi, A., Oyedele, L., Owolabi, H., Akinade, O., Bilal, M., Davila Delgado, J. M. & Akanbi, L. (2020). Deep Learning Models for Health and Safety Risk Prediction in Power Infrastructure Projects. *Risk analysis, 40*(10), 2019-2039. https://doi.org/10.1111/risa.13425

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., . . . Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced engineering informatics, 30*(3), 500-521. https://doi.org/10.1016/j.aei.2016.07.001

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature (London), 521*(7553), 436-444. https://doi.org/10.1038/nature14539

Kochovski, P. & Stankovski, V. (2021). Building applications for smart and safe construction with the DECENTER Fog Computing and Brokerage Platform. *Automation in construction, 124*, 103562. https://doi.org/10.1016/j.autcon.2021.103562

Sattari, F., Lefsrud, L., Kurian, D. & Macciotta, R. (2022). A theoretical framework for data-driven artificial intelligence decision making for enhancing the asset integrity management system in the oil & gas sector. *Journal of loss prevention in the process industries, 74*, 104648. https://doi.org/10.1016/j.jlp.2021.104648

Mirnezami, S. A., Mousavi, S. M. & Mohagheghi, V. (2020). An innovative interval type-2 fuzzy approach for multi-scenario multi-project cash flow evaluation considering TODIM and critical chain with an application to energy sector. *Neural computing & applications, 33*(7), 2263-2284. https://doi.org/10.1007/s00521-020-05095-z

Heo, S., Han, S., Shin, Y. & Na, S. (2021). Challenges of data refining process during the artificial intelligence development projects in the architecture, engineering and construction industry. *Applied sciences, 11*(22), 10919. https://doi.org/10.3390/app112210919

Chou, J., Lin, C., Pham, A. & Shao, J. (2015). Optimized artificial intelligence models for predicting project award price. *Automation in construction, 54*, 106-115. https://doi.org/10.1016/j.autcon.2015.02.006

Yaseen, Z. M., Ali, Z. H., Salih, S. Q. & Al-Ansari, N. (2020). Prediction of risk delay in construction projects using a hybrid artificial intelligence model. *Sustainability (Basel, Switzerland), 12*(4), 1514. https://doi.org/10.3390/su12041514

Chen, S. & He, J. (2012). *Research on cost management system of distribution network construction projects based on data mining*. https://doi.org/10.1109/CI-CED.2012.6508454

Choi, S., Lee, E. & Kim, J. (2021). The Engineering Machine-Learning Automation Platform (EMAP): A Big-Data-Driven AI Tool for Contractors' Sustainable Management Solutions for Plant Projects. *Sustainability (Basel, Switzerland), 13*(18), 10384. https://doi.org/10.3390/su131810384

Russell, S. J. & Norvig, P. k. (2014). *Artificial intelligence: A modern approach* (3. ed., Pearson new internat. ed.). Pearson.

Mullainathan, S. & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *The Journal of economic perspectives, 31*(2), 87-106. https://doi.org/10.1257/jep.31.2.87

MathWorks. (n.d). How the Genetic Algorithm works. Retrieved September 18, 2022, https://www.mathworks.com/help/gads/how-the-genetic-algorithm-works.html

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, SpaCy, and Keras*.

Lin, S., Shen, S., Zhou, A. & Xu, Y. (2021). Risk assessment and management of excavation system based on fuzzy set theory and machine learning methods. Automation in construction, 122, . https://doi.org/10.1016/j.autcon.2020.103490