DATA QUALITY ISSUES THAT HINDER THE IMPLEMENTATION OF ARTIFICIAL NEURAL NETWORK (ANN) FOR COST ESTIMATION OF CONSTRUCTION PROJECTS IN MALAYSIA

Alya Farhani Mohd Zammari[1,] Mohd Fairullazi Ayob[2*]

[1,2]Department of Quantity Surveying, Kulliyyah of Architecture and Environmental Design, International Islamic University Malaysia, Jalan Gombak, 53100, Kuala Lumpur, Malaysia.
**Corresponding author's email: fairullazi@iium.edu.my*

ABSTRACT

The Artificial Neural Network (ANN), which is one of the Artificial Intelligence (AI) tools, has been identified as a great technique to be used for construction cost estimation in the project. With the optimum quality of data input into the ANN model, it could produce an optimum and reliable cost estimation output. Nonetheless, the construction industry lacks the breadth and depth of data required as input into ANN. Though many online databases have been made available for data consumers, data quality problems remain unresolved. Thus, this study aims to identify data quality issues that can hinder the implementation of ANN for cost estimation of a construction project. Literature review and semi-structured interview were employed for the data collection of this research. The content analysis method was used to analyse the information obtained through the literature review. Meanwhile, the data collected from the semi-structured interview with nine (9) respondents was analysed using both content analysis and descriptive statistics analysis methods. The findings revealed six data quality issues that can hinder the ANN implementation for cost estimation of construction projects in Malaysia which are inaccurate data, outdated data, data access barriers, insufficient data, noise in training data, and data input degree of influence. Academically, this study contributes to the body of knowledge about the implementation of ANN for cost estimation of construction projects in Malaysia.

Keywords: Artificial Neural Network, Artificial Intelligence, Cost Estimation, Data Quality.

1.0 INTRODUCTION

Artificial neural network (ANN) is one of the Artificial Intelligence (AI) tools, part of the machine learning subfield that is used for mathematical optimisation (Adio-moses & Asaolu, 2016). ANN is an information-processing system or technique that is modelled after neural biology in the human brain (Patil, 2017). It can solve classification, prediction, and regression problems (Juszczyk, 2017). Therefore, concerning its prediction ability, the ANN technique can be utilised to perform cost estimation of the construction project. Primarily, the purpose of cost estimation is to yield an accurate prediction of early-stage cost estimates in comparison with the final cost of a construction project. Thus, the accuracy of estimation is a critical aspect of achieving success (Al-Tawal et al., 2021). ANN is data-driven and is considered sensitive to input data (Tayefeh Hashemi et al., 2020). This implies that to perform tasks such as prediction, ANN relies on data input. A reasonable amount of data must be input into the ANN model to be processed and produce a result or output, in which the result accuracy will rely on the quality of data input. Hence, to optimise the result or output produced by ANN, the quality of data input must be optimised.

According to the Construction Industry Development Board (CIDB) and the Royal Institute of Surveyors Malaysia (RISM), the Malaysian construction industry faces a lack of access, accuracy, breadth, and depth of industry data. Thus, the CIDB, in collaboration with the RISM, has established an online cost database known as Building Cost Information Services Malaysia (BCISM) as an initiative to provide open data to targeted data consumers in construction industry, such as quantity surveyors, contractors, engineers, researchers, etc. Besides, there are other online cost databases in

Malaysia, such as CIDB Construction Information for your Convenience (CONVINCE), Public Works Department (PWD) Rates Online, Juru Ukur Bahan Malaysia (J.U.B.M) and Arcadis and Quantity Surveyors Online. Nonetheless, although open-data popularity has increased, the problem with data quality remains unresolved (Nikiforova, 2020). This leads to this research aim to identify the data quality issues that can hinder the implementation of ANN for cost estimation of construction projects in Malaysia. Primary and secondary data collection were employed to achieve this aim.

2.0 THE REVIEW OF DATA INPUT FOR PROJECT COST ESTIMATION USING ARTIFICIAL NEURAL NETWORK (ANN)

2.1 Definition of Data Quality

High-quality data is important to avoid the consequences of data quality problems which can lead to financial losses. Daily, users encounter data because data exist everywhere. Hence, regardless of the level of knowledge in the information systems (IT) and data quality, each user must be able to verify the data quality (Nikiforova, 2020). Moreover, to verify the data quality, users must possess knowledge of what is defined as the quality of the data. Multiple researchers have defined and characterised the term 'data quality' in their research to provide a better understanding of the quality of data. Nikiforova (2020), stated that the concept of data quality is defined as:

> "…the suitability of the data for use case, emphasising its relative and dynamic nature, the context of which is determined by the data use and the requirements that depends on it and may change over time, that is determined by data gradual accumulation in the databases, and changing data quality requirements" (p. 391).

Meanwhile, other researchers defined data quality as "the ability of a data collection to meet user requirements" (Cappiello et al., 2004, p. 68), "measure of the agreement between the data views presented by information systems and that same data in the real world" (Orr, 1998, p. 3), "…data that is fit for the use of consumers" (Strong et al., 1997, p. 104) and "…data that are fit for use by data consumers" (Wang & Strong, 1996, p. 6). Fundamentally, data quality is the data that fulfils the requirements or criteria for the data consumers (Cappiello et al., 2004; Nikiforova, 2020; Strong et al., 1997; Wang & Strong, 1996). It is also an extent of agreement between data views offered by information systems and data in reality, which means that 100% data quality indicates the data views as perfect agreement with reality while 0% data quality indicates no agreement at all (Orr, 1998). A set of data must meet specific criteria or dimensions to be regarded as quality.

2.2 Data Input Required for Cost Estimation Using Artificial Neural Network (ANN)

ANN requires data input to process and generate the desired output. As mentioned previously in this article, ANN is a machine learning that is data-driven, and it is sensitive to input data (Tayefeh Hashemi et al., 2020). Weckman et al. (2010) and El-Sawalhi and Shehatto (2014) stated that the ANN model relies on historical data to recognise the pattern and generates output. Meanwhile, Ji et al. (2019) stated that ANN can form various types of models, and most of them rely on historical data. In terms of cost estimation of a construction project, data required as input into ANN is the cost and design parameters (Arafa & Alqedra, 2010; Chandanshive & Kambekar, 2019). Cost parameters refer to the total cost of a construction project for specific building elements. For example, using the ANN model, Roxas and Ongpeng (2014) used the total building structural cost from 30 building projects to estimate the cost of structural works. Dinh Cong and Nguyen Minh (2020) use total construction work expenses from available cost databases to estimate construction of school costs using ANN.

Table 1 shows the parameters of the building structural works used as data input to estimate the cost using the ANN model. The researchers that carried out the cost estimation of structural works using the ANN model have identified the most influential factors of cost estimation of building structure or known as the building parameter. "Influential factors of cost are elements which represent a building's features and impact its cost"(Ji et al., 2019, p. 4). Therefore, the most influential design parameters of a building or any building elements must be identified first to guarantee reliable cost estimates using ANN.

Table 1 Building structural works data input parameters

NO.	BUILDING STRUCTURE PARAMETERS	AUTHORS
1.	The total area of a building	(Arafa & Alqedra, 2010; Günaydin & Doğan, 2004; Roxas & Ongpeng, 2014)
2.	Area of ground floor	
3.	The ratio of the typical floor area to the total area of the building	
4.	The ratio of the ground floor area to the total area of the building	
5.	Number of floors/stories	
6.	Type of floor	
7.	Type of foundation	
8.	Number of elevators	
9.	Number of basements	
10.	Area of formworks	
11.	Weight of reinforcing steel	
12.	Volume of concrete	
13.	Console direction of the building	
14.	Location of the core of the building	

There are 14 influential parameters used as input for cost estimation of building structural works using ANN. Some input has more influence than others and affects the cost estimation result accuracy differently. For instance, Arafa and Alqedra (2010) carried out a sensitivity analysis and identified the most influential parameters of structural works cost estimation are ground floor area, number of stories, type of foundation, and number of elevators. Therefore, these four parameters are the most impactful and must be included as input when performing cost estimation of structural works using ANN. Overall, it is important to identify the most important and influential factors that affect the cost estimation of buildings or any elements, such as structural works, because they will be used as input into the ANN model along with the cost data from historical projects. The degree of influence of parameters will impact the accuracy of cost estimates using ANN. Other than concern with parameters, the historical project data must be documented properly to ease the process of data input into the ANN model.

2.3 Data Quality Issues That Hinder ANN Implementation for Cost Estimation of Construction Project

Data plays an enormous role in AI applications since it is the main factor that affects their performance. Many researchers have emphasised that machine learning, as part of AI applications,

will require a large amount of data. The larger the amount of data, the better the performance and efficiency, especially in terms of pattern recognition. Thus, ANN applies the same since it is part of AI and machine learning. It relies on the input data provided, processes the data, and produces output. Cost estimation using ANN will require data input, such as cost data from historical projects (Chandanshive & Kambekar, 2019). The data input is susceptible to a few issues linked to the data quality, as discussed below. Table 2 depicts the summary of the data quality issues that can hinder the ANN application for cost estimation of construction projects in Malaysia.

The first issue is regarding data quantity which is an insufficient amount of data. Adequate data is crucial for AI-based systems to learn from to perform more efficiently and intelligently (Jezova, 2018). One of the factors that affect the accuracy of cost estimates predicted by ANN is the size of the data (Al-Tawal et al., 2021). Ji et al. (2019), stated that sufficient data is crucial for cost estimation since most cost models are based on historical data. This means that to estimate the cost using the ANN model, the amount of data to be used as input must be adequate. An inadequate amount of data will affect the accuracy of cost estimates by ANN. For example, Richa et al. (2016) uses cost data from the schedule of rates for the past 23 years to estimate the cost of residential buildings using ANN and get result accuracy above 90%. Meanwhile, Chandanshive and Kambekar (2019) in their study find out that the accuracy of the ANN model increases with the data size. Hence, sufficient data input is important to estimate the cost using ANN so that the ANN model can perform better to recognise patterns and produce high-accuracy cost estimates.

The second issue is inaccurate data. An accurate database is important as it will produce the fastest and most effective cost estimation method (Al-Tawal et al., 2021). Data accuracy is important if high-accuracy cost estimates are the goal. Data accuracy is one of the important data quality attributes (Nikiforova, 2020). One example of inaccurate data is data that has errors such as typing errors. For example, cost estimation deals with figures from cost data as input. If an incorrect amount is entered, the result accuracy will not be reliable. For example, the cost per unit of concrete is RM170.00 per m^3, but it was input as RM 17.00 per m^3. Thus, the cost estimation amount predicted by ANN is inaccurate and wrong.

The third issue is data access barriers. Wang and Strong (1996) defined accessibility as “the extent to which data are available or easily and quickly retrievable (p. 32). Accessibility is associated with the ability to retrieve or access data from the information system. The accessibilities become issues when the access is limited for the sake of privacy and security. If any access barriers exist, data consumers will perceive them as accessibility problems (Strong et al., 1997). In Malaysia, many cost data producers publish by construction boards or companies such as CIDB CONVINCE. CIDB CONVINCE can be regarded as difficult to access since it requires registration, username, and password to access the data. Other than the ability to access the system, the user-friendliness of the system is also considered when it comes to data accessibility.

Fourth is the data currency issue. Currency is the degree to which the data's age is appropriate for the task at hand (Lee & Strong, 2003; Wang & Strong, 1996). For instance, outdated cost data. Cost data used to estimate the cost of a project must be recent and up to date in conjunction with the construction market, considering the fluctuating price. Using outdated or irrelevant cost data can lead to an underestimate or overestimation of cost estimation value by the ANN model. Hence, the output produced is not reliable in terms of accuracy.

Fifth issue is the noise in training data. According to Alharbi et al. (2021), noise in training data is also known as outliers which occur in datasets for a variety of reasons related to data collection, human errors, and the widespread use of suboptimal automated processes to compile large datasets. Noise is one of the key issues that are common with ANN modelling and it can degrade the performance of ANN model to produce reliable output (Abiodun et al., 2018; Alharbi et al., 2021). Although ANN is known as noise tolerance, handling the noise in data can solve many non-linear problem and provide optimum outcome or result.

All in all, the data quality issues discussed above, which are an insufficient amount of data, inaccurate data, data access barriers, data currency issues, and noise in training data could affect the accuracy of cost estimation value produced by ANN. Moreover, poor data quality will cause significant problems in developing an ANN model that is reliable for cost estimation prediction (Gudivada et al., 2017). This implies that it can hinder the implementation of ANN for cost estimation of construction projects. In addition, the quality of the result produced by ANN will highly depend on the data quality that is input into the ANN model. It is important to address these issues to produce a reliable and high-quality cost estimate. There is limited literature that discusses the data issues of AI specifically for cost estimation. Hence, the general data issues regarding AI data were discussed, and the understanding was applied to cost estimation practice using ANN.

Table 2 Data quality issues

NO.	DATA QUALITY ISSUES	AUTHORS
1.	Inaccurate data	CNIL (2017), Francisco et al. (2017), Ji et al. (2019)
2.	Insufficient amount of data	CNIL (2017), FRA (2019), Ji et al. (2019)
3.	Data access barriers	Francisco et al. (2017), Strong et al. (2017), Al-Tawal et al. (2021)
4.	Data currency issue	Wang & Strong (1996), Lee & Strong (2003), Francisco et al. (2017)
5.	Noise in training data	Alharbi et al. (2021), Abiodun et al. (2021)

3.0 METHODOLOGY

Figure 1 presents the overall process involved which is the essence of the methodology adopted in this research. Research methodological choice, data collection approaches and data analysis techniques will be explained further in this section of this paper.

3.1 Research Methodological Choice

The qualitative approach was adopted in carrying out this research and achieving the aim and objectives. This approach was employed because the research nature is subjective, and it is required to explore the opinion, views, and perceptions of respondents on the quality of data for cost estimation using the ANN technique. Furthermore, the qualitative approach was selected because this research involved a limited number of respondents who has an ANN background. Only people who possess knowledge, skills, or experience in ANN can become the respondents and share their knowledge and opinions through both open-ended and close-ended questions. Besides, the qualitative approach was adopted to gain a degree of understanding of the topic that a closed-question

survey cannot provide.

3.2 Data Collection Approach

This research employed both data collection approaches to gather the required information for achieving the aims and objectives. Secondary data collection was done first before the primary data collection was conducted.

3.2.1 Secondary data collection

Various sources were referred to during the literature review process, which are journal articles, books, conference proceedings, online reading materials, dissertations, and government publications. The literature review process was essential because it helped to understand the research topics deeper and find the gaps in previous literature studies. Not only the researcher but it also aided the reader in better understanding what the study is all about. Moreover, its findings served as the basis for constructing questions for primary data collection.

3.2.1 Primary data collection

This research opted for a semi-structured interview to collect data as the primary source to fulfil the aim of this research. Semi-structured interviews involve asking both open-ended and close-ended questions to interviewees without having a specific sequence or schedule (Naoum, 2007). It allows the interviewer to gain new insights and interviewees to raise unexpected questions (Jilcha Sileyew, 2020). Considering this research is related to ANN practice, which is still rudimentary in Malaysian construction, semi-structured interviews with experts are the most appropriate method to obtain primary sources. Both open-ended and close-ended questions were asked to gain the knowledge and opinions of the respondents regarding this topic. Moreover, although some key questions and answers were predetermined, it also allowed both interviewer and interviewees to raise other questions or share opinions regarding the topics which can contribute to more knowledge.

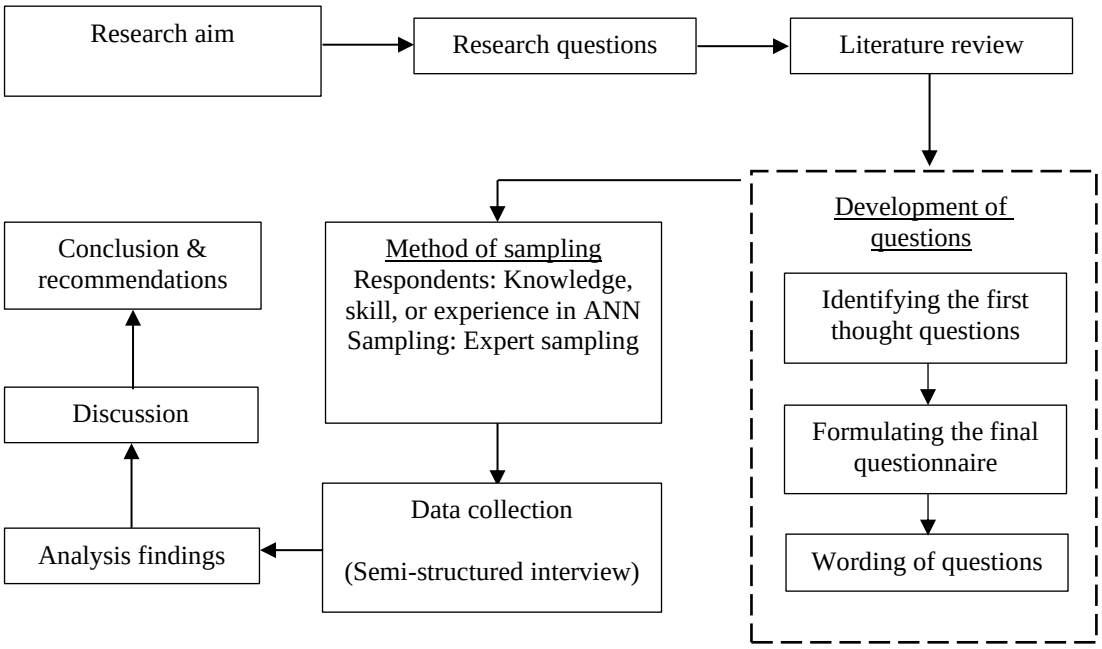

Fig. 1: Overview of the research process (Edited from Naoum, 2007)

3.2.3 Sample selection

This study used expert sampling, which involves those who can elucidate and demonstrate using their experience or those who are specialists in a particular field (Etikan, 2017). For this research, the criteria of the targeted respondents are those who have knowledge, skill, or experience in ANN. The respondents could be anyone from built environment backgrounds, such as architects, quantity surveyors, engineers, etc. Apart from that, those who are not from built environment backgrounds can also participate as respondents if they have knowledge, skill, or experience with ANN. Apart from expert sampling, the researcher also adopted snowball sampling, which is a selection of respondents by using networks (Etikan, 2017). This method was intended to find more experts in ANN since the number of practitioners is still limited. First, expert sampling was done by choosing only those who have knowledge, skill, or experience in ANN. Then, snowball sampling was done through the selected experts that are qualified to participate as respondents of this research. In the interview questions, the selected experts were asked to recommend any person that has knowledge, skill, or experience in ANN and is willing to participate in this study.

3.2.4 Data collection

The semi-structured interview was the primary data collection technique selected by the researcher. The respondents for this research were obtained mainly through a literature search, projects, and recommendations from respondents or also known as snowball sampling. The researcher has selected respondents who have knowledge, skills, or experience regarding ANN. There are 55 potential respondents identified by the researcher through multiple platforms, which are email and social media such as LinkedIn, WhatsApp, Telegram and Facebook. Eventually, a total of nine respondents agreed to participate in this research to make a contribution as experts in ANN. The following Table 3 shows the details of each respondent:

Table 3 Respondents' background

RESPONDENT	DESIGNATION/ SPECIALISATION	ORGANISATION	EXPERIENCE IN THE CONSTRUCTION INDUSTRY
1	**Mechanical Engineer** (Artificial Intelligence/ Robotic and Automation)	Malaysia Airlines Berhad	11-20 years
2	**Machine Learning Engineer** (Machine Learning/ Data Science)	*Not provided*	Less than 3 years
3	**Electronic Engineer** (Artificial Intelligence)	*Not provided*	Less than 3 years
4	**Mechanical Engineer** (Programmer/ Data Analysis)	Sony EMCS (Malaysia) Sdn. Bhd.	Less than 3 years
5	**Engineer** (Data Science/ Machine Learning)	*Not provided*	3-5 years
6	**Lecturer** (Mechanical and Aerospace Engineering)	International Islamic University Malaysia (IIUM)	11-20 years
7	**Lecturer** (Biotechnology Engineering/ Information and Communication Technology (ICT)/ Artificial Intelligence)	International Islamic Universiti Malaysia (IIUM)	6-10 years
8	**Lecturer** (Civil Engineering)	Universiti Teknologi Mara (UiTM)	11-20 years
9	**Data Scientist**	*Not provided*	Less than 3 years

3.3 Data Analysis Techniques

Data analysis is important to summarise and organise data in a meaningful and effective way (Naoum, 2007). It is also to present the findings of the research and its correlation to the research aim and objectives. Since the questions comprised both open-ended and close-ended, the researchers employed content analysis and descriptive statistics analysis to analyse the data obtained from the

respondents. Open-ended questions are analysed using content analysis, while close-ended questions are analysed using descriptive statistical analysis through Microsoft Excel 2019. Descriptive statistics analysis includes the frequency and percentage distribution, mean and standard deviation. Frequency and percentage are used for multiple choice questions, while mean and standard deviation are used to analyse questions involving the Likert Scale.

3.3.1 Mean ranking analysis

Mean ranking was used to obtain the average value of data collected from respondents. The total of data values was added and divided by the dataset number (number of respondents) to calculate the mean. The following Figure 2 shows the formula to calculate the ungrouped mean.

$$\bar{x} = \frac{\sum x}{N}$$

$$\sum x = the\ sum\ of\ X$$

N = number of data

Fig. 2: Formula to calculate the ungrouped mean
Source: https://spmaddmaths.blog.onlinetuition.com.my/2013/10/7-1a-mean.html

3.3.2 Standard deviation

Standard deviation is used to measure the dispersion of the dataset concerning its mean. In this research, the standard deviation value for the dataset will indicate the respondents' degree of consensus on the related subject. For instance, the consensus on the degree of agreement on the data quality issues and mitigation strategies for the improvement of ANN practice for project cost estimation. The following Figure 3 shows the formula to calculate the standard deviation:

$$\boldsymbol{\partial} = \sqrt{\frac{\sum(x - mean)^2}{n}}$$

x= set of numbers
Mean= average of the set of numbers
n= size of the set
$\partial = the\ standard\ deviation$

Fig. 3: The formula to calculate the standard deviation
Source: https://www.basic-mathematics.com/standard-deviation-formula.html

Generally, a standard deviation value that is more than 0.00 indicates that there is a disagreement between the respondents. A low standard deviation value means that the data is clustered around the mean, while a high standard deviation value means that the data is more spread out from the mean. The lower the standard deviation value and closer to 0.00, the higher the consensus among the respondents. The higher the standard deviation value, the lower the consensus among respondents on related subjects. For the result analysis of this article, the level of consensus achieved from the standard deviation value is presented in Table 4.

Table 4 Level of consensus achieved from standard deviation value

Standard Deviation (SD)	Level of consensus achieved
$0 \leq X < 1$	High level of consensus
$1 \leq X < 1.5$	Reasonable/ fair level of consensus
$1.5 \leq X < 2$	Low level of consensus
$2 \leq X$	No consensus

Source: Grobbelaar (2007); as cited in Bidi & Ayob (2015)

4.0 RESULTS

4.1 Data Quality Issues That Can Hinder the Implementation of ANN Practice in The Construction Project Cost Estimation

This section reports and presents the data collected from the primary data collection, which is the semi-structured interview conducted with nine respondents. The respondents were asked to rate the degree of agreement on the data quality issues that can hinder the implementation of ANN practice in project cost estimation. The researcher highlights four issues with specific references to data quality that were obtained through a literature review.

Table 5 presents the mean and standard deviation values for each issue. The first rank is shared between two issues that recorded the same value of the mean. 'Inaccurate data that involve many errors' and 'outdated data either from historical projects or published cost data' have recorded a 4.33 mean value. Both issues recorded a 0.71 standard deviation value indicating a high level of consensus among respondents on the average score. The second highest is the 'access barriers to retrieve data from a known source such as online database', which recorded a 4.22 mean value. The standard deviation value is 0.97 indicating a high level of consensus among respondents. Lastly is the 'insufficient amount of data to be used as input for cost estimation' that recorded a 4.00 mean value. The standard deviation value achieved is 1.12 which indicates a fair level of consensus among respondents.

Table 5 Mean and standard deviation value for data quality issues

NO.	DATA QUALITY ISSUES IN ANN PRACTICE	MEAN	SD
1.	Inaccurate data that involve many errors	4.33	0.71
	Outdated data, either from historical projects or published cost data	4.33	0.71
2.	Access barriers to retrieving data from a known source, such as an online database	4.22	0.97
3.	Insufficient amount of data to be used as input for cost estimation	4.00	1.12

Other than the data quality issues provided in Table 5, the respondents have given their opinion on other data quality issues that can hinder the implementation of ANN practice in project cost estimation. The list of the issues is provided below:

1) Noise in training data
2) Degree of influence of data input parameters

5.0 DISCUSSIONS

The result analysis from this study found that the data quality issues that are most impactful towards the implementation of ANN for cost estimation are the "inaccurate data that involve many errors" and "outdated data either from historical projects or published cost data" where both issues are ranked the first place. The second place is "access barriers to retrieve data from a known source such as an online database" while on third place is "insufficient amount of data to be used as input for cost estimation". This result is based on the respondents' degree of agreement on the data quality issues obtained by the researcher through literature reviews that, in their opinion, can hinder the implementation of ANN for cost estimation of construction projects.

Since the issues addressed in this research are specific to data quality, the researcher expected the result would prioritise the issues directly linked to the four data quality attributes used as a framework for this research: availability, accessibility, currency, and reliability. As an unanticipated finding from this result, one of the issues ranked first concerns data accuracy. A study by Cichy and Rass (2019), has included data accuracy as one of the most important data quality attributes but is not specific to any type of data. Moreover, Al-Tawal et al. (2021) emphasise the importance of accurate databases since they will produce the most effective cost estimation method. Despite this, the researcher believes that issues concerning data accuracy are critical and need to be addressed because they will surely impact the cost estimation result produced by ANN.

"Outdated data from historical projects or published cost data" reflects the issues regarding data currency. In contrast to the word "outdated", data must be current or recent. Most of the respondents agreed that data currency is the critical factor that can hinder ANN implementation for cost estimation. A set of databases, especially for cost data, must be updated to apply to current market trends and situations. For example, the agencies that have published online cost databases must frequently update their data in the system regularly, either monthly, quarterly, or yearly. Moreover, Liu et al. (2021) stated that to achieve sustainable data updating and good data quality, data standards must be updated and an automated information-gathering system implemented. The data quality issues that fall on the second rank, which are the access barriers to retrieving data from a known source, such as online cost databases, reflect the data quality regarding accessibility. This issue recorded a slightly higher standard deviation value than the previous data, implying that the consensus among respondents is lower. This is affected by the score given by one of the respondents, which deviates from the mean score. Respondent 3 rated a score of 2, indicating disagreement for this issue as one of the factors that hinder the ANN practice for cost estimation. Based on the background of the specified respondent, the researcher assumes that the little experience in both the construction industry and ANN practice makes the respondents rarely encounter this issue and fail to see the correlation between this issue and the hindrance of ANN practice for cost estimation. One example of data accessibility issues is the access barriers through online cost databases that require personal registration or login into the website. Wang & Strong (1996) stated that if there is any barrier, the data consumer will regard it as a data access problem. Moreover, if the retrieval of data is difficult, it is challenging to obtain data as input into the ANN model.

Insufficient amount of data to be used as input for cost estimation; the issues on the third rank represent the data size or quantity. Of most respondents that rated 3, 4, and 5 for this issue, respondent 3 rated score 2 for this issue, resulting in only a fair level of consensus among respondents. Respondent 3, which has an engineering and artificial intelligence background, disagrees that

insufficient data can hinder the implementation of ANN for cost estimation. This could be due to the respondent's lack of experience since this respondent only has less than three years of experience in the construction industry and ANN practice. On the other hand, respondents 2, 5, and respondent 8 supported this issue not only by rating 4, 5, and 5 scores, respectively, but they also opined that inadequate data could hinder ANN practice for cost estimation. Furthermore, they emphasise that the government and related companies must take initiatives to share more related data in the industry since data drives the development of AI and ANN. From the researcher's point of view, this initiative from the government will benefit them because the development and implementation of AI and ANN is one of the enablers of the fourth industrial revolution. Next, a few of the respondents have provided other data quality issues that, in their opinion, can hinder the practice of ANN for cost estimation of construction projects.

Respondent 4 opined that noise training data could be one of the issues that hinder ANN practice for cost estimation. Similar to the study by Abiodun et al. (2018), one of the key issues in the ANN model is noise. Moreover, Matel et al. (2019), stated that the noise in training data is a typical cause of variation in the ANN final model. According to Alharbi et al. (2021), noise in training data refers to the outliers that occur in data sets for a wide range of reasons, including data collection, human errors, and the frequent use of poor automated procedures to construct huge datasets. One example of noise is the data with meaningless information that cannot be interpreted and analysed appropriately.

Respondent 8 stated that one of the issues is the degree of influence of data as input parameters in the ANN model. He underlined that the data are not researched well enough to determine whether they are the most important data to include in databases. This issue seems to be consistent with other research which found that determining the most influential parameters as data input to perform cost estimation using ANN is crucial (Arafa & Alqedra, 2010). This implies that the degree of data influence affects the performance of the ANN model in cost estimation.

All in all, the researcher has identified six data quality issues that can hinder the implementation of ANN for cost estimation of a construction project through both primary and secondary data collection, which are inaccurate data, outdated data, data access barriers, an insufficient amount of data, noise training data, degree of influence of data input parameters as well as long period to train ANN model. From the discussion above, it is apparent that the data quality issues are impactful not only towards the quality of cost estimate results produced by ANN but also towards the hindrance of ANN practice for cost estimation of construction projects.

6.0 CONCLUSION

This paper presents the outcomes of this study on the data quality issues that can affect the cost estimation accuracy as well as hinder the ANN practice in construction cost estimation. The outcome of this study has established six data quality issues identified, which are inaccurate data that involve many errors, outdated data either from historical projects or published cost data, access barriers to retrieving data from a known source such as an online database, an insufficient amount of data to be used as input for cost estimation noise in training data and degree of influence of data input parameters. As a suggestion, it is recommended that the producer of the database be alert to the occurrence that is perceived as an issue by the data consumer. Aside from that, cost estimators who will use ANN as a cost estimation technique can assess the quality of the data before performing cost

estimation using ANN to avoid incorrect and unreliable cost estimation results. This research helps to create awareness among construction professionals regarding the cost estimation technique, which is the ANN, one of the crucial integrated technologies in IR4.0. Construction professionals, especially quantity surveyors, can analyse the gap between their current practice for cost estimation and the implementation of the ANN technique. This research also contributes to the body of knowledge associated with Artificial Intelligence (AI) as Artificial Neural Networks (ANN) are part of AI. Furthermore, no previous research has been conducted on cost estimation using ANN techniques, specifically data quality.

ACKNOWLEDGMENTS

This paper follows the other paper below that has been presented elsewhere by the first author from nearly 9 months of research under the second author supervision:

Alya, Farhani (2022). Artificial Neural Network (ANN) Application for cost estimation of construction projects in Malaysia: A study on the quality of data, [Unpublished Degree's Dissertation]. International Islamic University Malaysia.

Essentially, the outcome of this research has been presented elsewhere by the authors in the form of poster and video and won Gold Award for the Best Poster and Video Presentation in the Invention and Innovation Research Competition, as shown follows:

Alya, F., Ayob, M. F., (2022). Artificial Neural Network (ANN) Application for cost estimation of construction projects in Malaysia: A study on the quality of data, International Islamic University Malaysia. RISM Invention and Innovation Research Competition, 2022

REFERENCES

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. E., & Arshad, H. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon*, *4*(11), 1–41. https://doi.org/10.1016/j.heliyon.2018.e00938

Adio-moses, D., & Asaolu, O. S. (2016). Artificial intelligence for sustainable development of intelligent building. *Research Gate*, *9* (February), 1–11.
https://www.researchgate.net/publication/299437528%0AARTIFICIAL

Alharbi, F., El Hindi, K., Al Ahmadi, S., & Alsalamn, H. (2021). Convolutional Neural Network-Based Discriminator for Outlier Detection. *Computational Intelligence and Neuroscience*, *2021*. https://doi.org/10.1155/2021/8811147

Al-Tawal, D. R., Arafah, M., & Sweis, G. J. (2021). A model utilising the artificial neural network in cost estimation of construction projects in Jordan. *Engineering, Construction and Architectural Management*, *28*(9), 2466–2488. https://doi.org/10.1108/ECAM-06-2020-0402

Arafa, M., & Alqedra, M. (2010). Early Stage Cost Estimation of Buildings Construction Projects using Artificial Neural Networks. *Journal of Artificial Intelligence*, *4*(1), 63–75. https://doi.org/10.3923/jai.2011.63.75

Bidi, N. K., & Ayob, M. F. (2015). Investigation of Quality of Cost Data for Life Cycle Cost Analysis in. *14th Management in Construction Researchers' Association (MiCRA 2015)*, *1*, 1–15. http://irep.iium.edu.my/45835/

Cappiello, C., Francalanci, C., & Pernici, B. (2004). Data quality assessment from the user's perspective. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 68–73. https://doi.org/10.1145/1012453.1012465

Chandanshive, V. B., & Kambekar, A. R. (2019). Estimation of Building Construction Cost Using

Artificial Neural Networks. *Journal of Soft Computing in Civil Engineering*, *3*(1), 91–107. https://doi.org/10.22115/SCCE.2019.173862.1098

Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, *7*, 24634–24648. https://doi.org/10.1109/ACCESS.2019.2899751

Dinh Cong, T., & Nguyen Minh, Q. (2020). Estimating the construction schools cost in Ho Chi Minh City using Artificial Neural Network. *IOP Conference Series: Materials Science and Engineering*, *869*(6), 1–7. https://doi.org/10.1088/1757-899X/869/6/062014

El-Sawalhi, N. I., & Shehatto, O. (2014). A Neural Network Model for Building Construction Projects Cost Estimating. *Journal of Construction Engineering and Project Management*, *4*(4), 9–16. https://doi.org/10.6106/jcepm.2014.4.4.009

Etikan, I. (2017). Sampling and Sampling Methods. *Biometrics & Biostatistics International Journal*, *5*(6), 215–217. https://doi.org/10.15406/bbij.2017.05.00149

FRA. (2019). Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. *FRA Focus*, 18.

Francisco, M. M. C., Alves-Souza, S. N., Campos, E. G. L., & De Souza, L. S. (2017). Total data quality management and total information quality management applied to costumer relationship management. *ACM International Conference Proceeding Series*, *June 2018*, 40–45. https://doi.org/10.1145/3149572.3149575

Gudivada, V. N., Ding, J., & Apon, A. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations Research Topic "Data Quality for Big Data and Machine Learning" in Frontiers in Big Data View project Web Search Engines View project Data Quality. *International Journal on Advances in Software (2017)*, *10.1*(July), 1–20. https://www.researchgate.net/publication/318432363

Günaydin, H. M., & Doğan, S. Z. (2004). A neural network approach for early cost estimation of structural systems of buildings. *International Journal of Project Management*, *22*(7), 595–602. https://doi.org/10.1016/j.ijproman.2004.04.002

Jezova, D. (2018). Artificial intelligence and privacy. *Human Rights - From Reality to the Virtual World*, 226–237. https://doi.org/10.13166/wsge//xuxl1897

Ji, S. H., Ahn, J., Lee, H. S., & Han, K. (2019). Cost Estimation Model Using Modified Parameters for Construction Projects. *Advances in Civil Engineering*, *2019*, 1–10. https://doi.org/10.1155/2019/8290935

Jilcha Sileyew, K. (2020). Research Design and Methodology. *Cyberspace*, 1–12. https://doi.org/10.5772/intechopen.85731

Juszczyk, M. (2017). The Challenges of Nonparametric Cost Estimation of Construction Works with the use of Artificial Intelligence Tools. *Procedia Engineering*, *196*(June), 415–422. https://doi.org/10.1016/j.proeng.2017.07.218

Lee, Y. W., & Strong, D. M. (2003). Knowing-Why About Data Processes and Data Quality. *Journal of Management Information Systems*, *20*(3), 13–39. https://doi.org/10.1080/07421222.2003.11045775

Liu, S., Chang, R., Zuo, J., Webber, R. J., Xiong, F., & Dong, N. (2021). Application of artificial neural networks in construction management: Current status and future directions. In *Applied Sciences (Switzerland)* (Vol. 11, Issue 20). https://doi.org/10.3390/app11209616

Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T., & Voordijk, H. (2019). An artificial neural network approach for cost estimation of engineering services. *International Journal of Construction Management*, *0*(0), 1–14. https://doi.org/10.1080/15623599.2019.1692400

Naoum, S. G. (2007). Dissertation research and writing for construction students. In *Elsevier* (Vol. 7, Issue 1).

https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download%0Ahttp://www.econ.upf.edu/~reynal/Civil wars_12December2010.pdf%0Ahttps://think-asia.org/handle/11540/8282%0Ahttps://www.jstor.org/stable/41857625

Nikiforova, A. (2020). Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. In *Baltic Journal of Modern Computing* (Vol. 8, Issue 3). https://doi.org/10.22364/BJMC.2020.8.3.02

Orr, K. (1998). Data Quality and Systems Theory. In *Communications of the ACM* (Vol. 41, Issue 2, pp. 66–71). https://doi.org/10.1145/269012.269023

Patil, S. (2017). Artificial Intelligence for the Prediction of Safe Work Behavior in Construction. *International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)*, *9*(5), 1418–1423. http://www.ijirset.com/upload/2017/january/162_Suraj survey_IEEE.pdf

Richa, Y., Monika, V., Vivekanand, V., & Sanket, A. (2016). Cost Estimation Model (Cem) for Residential Building using Artificial Neural Network. *International Journal of Engineering Research And*, *V5*(01), 430–432. https://doi.org/10.17577/ijertv5is010431

Roxas, C. L. C., & Ongpeng, J. M. C. (2014). An Artificial Neural Network Approach to Structural Cost Estimation of Building Projects in the Philippines. *DLSU Research Congress*, 1–8. https://www.dlsu.edu.ph/wp-content/uploads/pdf/conferences/research-congress-proceedings/2014/SEE/SEE-I-005-FT.pdf%0Ahttp://www.dlsu.edu.ph/conferences/dlsu_research_congress/2014/_pdf/proceedings/SEE-I-005-FT.pdf

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110. https://doi.org/10.1145/253769.253804

Tayefeh Hashemi, S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: a systematic review on machine learning techniques. *SN Applied Sciences*, *2*(10), 1–27. https://doi.org/10.1007/s42452-020-03497-1

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. http://www.jstor.org/stable/40398176.

Weckman, G. R., Paschold, H. W., Dowler, J. D., Whiting, H. S., & Young, W. A. (2010). Using Neural Networks with Limited Data to Estimate Manufacturing Cost. *Journal of Industrial and Systems Engineering*, *3*(4), 257–274. http://www.jise.ir/article_4015.html