Comprehensive Approach to Assure the Quality of MSME Data in Indonesia: A Framework Proposal

Mujiono Sadikin[1], Adi Trisnojuwono[2], Rudiansyah[2] and Toni Widodo[3]
[1]*Faculty of Computer Science, Universitas Bhayangkara Jakarta Raya, Kota Bekasi, Indonesia*
[2]*Deputy of Enterpreneurship, Ministry of Coperation and Small Medium Enterprise, Jakarta, Indonesia*
[3]*Mitreka Solusi Indonesia, Jakarta, Indonesia*

Keywords: MSME, Data Quality, Data Cleansing, Attribute Domain Constraint, Relation Integrity Rule.

Abstract: As mandated by the laws and regulations that have been released, the Government of Indonesia has decided that Cooperation and MSME empowerment policies must be determined based on accurate Cooperation and MSME data profiles. Therefore, the Government of Indonesia, in this case the Ministry of Cooperation SMEs, executes a complete data collection program of Cooperation and MSME profile. Due to the characteristics and constraints of data collection, many risks must be mitigated. The main risk identified in this program is the possibility of reduced data quality for Cooperation and MSME caused by some factors. This paper presents a proposed comprehensive framework to ensure the quality of Cooperation and MSME data based on Khan's data quality criteria previously defined. The aim of the proposed framework is to prevent, detect, repair, and recover dirty data to achieve the required minimum standards of data quality. The proposed framework covers all stages and aspects of the data collection process. In the data cleaning and correction stage, we investigate on many techniques namely rule based, selection based, and machine learning based. In the initial validation of the framework presented in this paper, the results of several data cleansing methods applied are discussed.

1 INTRODUCTION

Due to the importance of data quality in all human live field, the management of data quality studies has got more attention from many researchers and practitioners as well (Liu et al., 2018). Data quality is a crucial issue must be tackled. Many researchers have published the results of their studies to address data quality issues specifically in certain areas. Several publications on data quality include: in the health sector (Dziadkowiec et al., 2016), online sales and marketing sector (Arndt et al., 2022), company registration cases in several countries (Nikiforova, 2020), environment monitoring (Zhang and Thorburn, 2022), concrete analysis in infrastructure sector (Ouyang et al., 2021), and manufacture industry (Martinez-Luengo et al., 2019).

Many approaches to assure the data quality fulfill the requested standard have been proposed. Lie et al summarised those approaches as published in (Liu et al., 2018). In their studies authors also investigate the application of some techniques and algorithms in data cleansing that are performed in various analysis stages according to data type. Based on the three previous general methods, the authors make improvements by adding visualization block. Despite the advantages of the proposed framework, the framework has not been validated to the real data case. A data quality improvement technique that combines Kahn's work and SPSS query syntax was proposed by Dziadkowiec et al (Dziadkowiec et al., 2016). By utilizing the Kahn Data Quality Framework, the authors claim that the quality of the resulting data can be better. In this paper, we adapt Kahn framework as part of the MSME data quality assurance framework by elaborating the possibility of implementing ML and DL approaches.

Another approach regarding quality data improvement is proposed by (Arndt et al., 2022). In their study, authors use filters to improve the quality of online market analysis data. The data used in this experiment was provided through a crowdsourcing service, i.e. Mturk. The aim of this methodology is to compare each of the four filter categories based on the data sources. Those four used filterers are direct selection, direct accuracy, statistical selection, and statistical accuracy.

User oriented data-driven methods is proposed by

Sadikin, M., Trisnojuwono, A., Rudiansyah, . and Widodo, T.
Comprehensive Approach to Assure the Quality of MSME Data in Indonesia: A Framework Proposal.
DOI: 10.5220/0012445700003848

In *Proceedings of the 3rd International Conference on Advanced Information Scientific Development (ICAISD 2023)*, pages 154-161
ISBN: 978-989-758-678-1

(Nikiforova, 2020). This approach contains three components, namely data objects, data quality specifications, and data quality measurement processes. The proposed model is applied to open data set of register companies in several countries such as Latvia, Norway, England, and Estonia. The authors claim that the proposed method is able to overcome the weaknesses left by the previous approach. Despite the proposed method is quite comprehensive, it only handles data that meets certain criteria such as: complete, free of ambiguous values, and correct. This such constraints present that the data quality issues continuously provide new challenges must be addressed.

Another aspect that must be considered to ensure data quality is the proper handling of master data (Prokhorov and Kolesnik, 2018). The proposed master data model management system consists of three activities: consolidation, harmonization, and management. In the consolidation stage, improvements are applied to the data structure and data collection. In the harmonization stage, alignment, normalization and classification are carried out, whereas in the management action stage that must be carried out is centralization and management.

Data quality assurance is more challenging when data collection processes are almost real time, such as highfrequency water quality monitoring systems as investigated by Zhang et al (Zhang and Thorburn, 2022). Many factors influence the decreasing of the real time data quality such as network problems, device malfunction, device replacement etc. To overcome the missing values of real time data set that will affect the quality of the information provided, the authors developed a cloud-based system that combines several techniques and advance algorithms to perform the missing values imputation. Several imputation techniques used in the system such as Mean Imputation, LOCF, Linear Imputation, EM, MICE, Dual-SSIM and M-RNN. Among these techniques, by overall Dual-SSIM provides the best performance when it applied to nitrate and water temperature data. Despite the methods is powerful in handling the real time data, however it does not handle integrity between data or integrity between attributes / attributes relations.

Real world data is mostly dirty due to errors found in data sets. In many cases, the actual data set is inconsistent, contains missing values, lacks of integrity, ambiguous, and contains outliers. Therefore data cleaning is not only the main task, but also the most important part of data management since data quality determines the quality of the information produced (Ridzuan and Zainon, 2019). The handling of data cleansing cases requires a different approach. Ouyang et al, as presented in (Ouyang et al., 2021), used a ML-based ensemble approach to detect outliers in a concrete measurement regression data set. The technique used in this case is the ANN-based model compared to KNN, LOF, COF, OCSVM, IFOREST, ABOD, and SOS. The approach used in the study is the best algorithm selection approach with forward and backward techniques. Based on the experimental results, ANN gives the best results in detecting outliers in the regression data used.

Noisy data that contributes to unreasonable decision-making also occurs in the energy manufacturing industry such as the offshore wind turbine structures health data collected through the SCADA monitoring system. To overcome the problem, (Martinez-Luengo et al., 2019) proposes a method based on ANN techniques to improve data quality through automation of data cleaning processes. The proposed framework consists of two steps: data noise checking and removal, and missing data imputation. This research was conducted to improve the quality of fatigue assessment on turbines which are thought to be heavily influenced by the quality of monitoring data generated by SCADA sensors. Therefore, in his research the authors compared the quality of data without cleaning with the quality of data with cleaning using the proposed method. From the experimental results, the authors concluded that the quality of the data after cleaning proved to be better.

In their review of big data cleansing, (Ridzuan and Zainon, 2019) summarize some methods can be applied to the purposes. Those methods are developed based on various techniques such as rule-based, ML-based, and knowledge-based. However, those existing methods contain some limitations when it deals with dirty data.

The complete data collection of Cooperations and SMEs conducted by the Ministry of SMEs is a unique data collection model. The uniqueness, complexity, and problems are found in all components, including area coverage, individual data targets, data collection model which is performed manually, various skill and knowledge of data collection officers, the complexity of data entry forms, short time allocated, and the project management as well. In terms of area coverage, the project covers more 240 district, 34 provinces crossover Indonesia Country with various topographies and land contours. The data collection is carried out manually by more than 1.000 enumerators. As other models of real data collection, data quality is also a major issue that must be resolved before further use of the data. Due to this uniqueness and exclusiveness, to the author's knowledge, there is no model/approach that can deal with this data quality

standard problem comprehensively. Therefore, this paper presents a proposed framework for cleaning the data obtained by this kind of collection model. In general, the proposed framework consists of components/steps specific to the case and a combination of existing/published approaches. The proposed framework also consists of various approaches according to the case found.

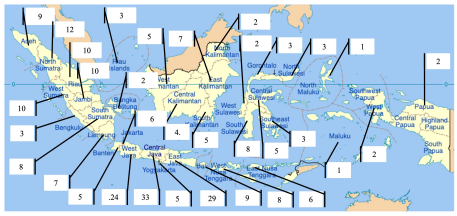

Figure 1: The distribution of City/Regency of data collection area.

2 RESEARCH METHOD

2.1 Overview of MSME Data Collection Program

This paper presents a framework proposal and case study of data cleansing in the Complete Data Collection performed by Ministry of Cooperation and MSE. This section discuss the overview of the program i.e. data collection object, its attributes, and the issues have to be resolved.

The objects of the data collection are the business actors of cooperation , micro, small and medium enterprises. In accordance with the mandate of the laws and regulations, the provision of single data for this business group is the duty and authority of the Ministry of Cooperations and SMEs (Indonesia, 2021). To carry out the duties and responsibilities of developing a single data for Cooperations and MSMEs, in the year of 2022 data collection is performed for more than 9 million individual Cooperation and MSMEs data by name by address. In terms of business location, micro and small business actors are very dynamic and vary. In the year stage, the data collection is limited to business actors who occupy a permanent business location.

For data collection purposes, each individual data is identified by 237 attributes which are divided into 15 blocks (groups) of attributes. Those attribute data blocks are: identity of place of business, identity of business characteristics, identity of business actors, business licenses, awards received, raw materials, production, labor, production processes, partnerships, financial business, coaching ever received, additional notes , and information from the registrar. The number of attributes in each block varies from 10 to 53.

The data collection area covers 34 Provinces and 240 Regencies/Cities/Districts in Indonesia. The distribution of districts/regencies/cities for data collection in all Indonesia Provinces is presented in Fig. 1. For each province, based on considerations of ease of transportation access and adequacy of infrastructure, there were chosen between 1 (one) to 33 regencies/districts/cities of each province. The characteristics of regencies/cities in the Indonesian territory widely vary in terms of topography, availability of infrastructure, social economy, and level of education of business actors. In the first stage of this data collection, districts/cities were selected based on considerations of the populationof business actors, affordability, and availability of data communication infrastructure.

Data collection was carried out by enumerators. The number of enumerators varies from one district/city to another, from 1 to. 1.709 enumerators. The parameters for the number of enumerators are based on consideration of the estimated number of business actors in the district/city concerned, the size of the coverage area, and the degree of access difficulty. This data collection also requires coordination across ministries/agencies both at the central and regional levels considering that the development of Cooperation and MSME business actors is cross-sector. Complete data collection is done by utilizing web-based applications and mobile apps that are accessed using the internet. This application is specifically prepared for -one of them - the implementation of this data collection.

Given the complexity of the complete Cooperation and MSME data collection program both in terms of enumerators, data collection areas, and the number of parties involved, the potential for data collection errors is quite large. Enumerators (data collectors) come from various educational and cultural backgrounds. Even though training and socialization have been carried out regarding the process and how to use the application for complete data collection, due to different levels of expertise, the potential for error entry by enumerators still exists. Each district/city is given a quota/target for a certain amount of individual data which is then distributed to enumerators. The minimum target that must be achieved also has the potential to cause enumerators to make mistakes in data collection, either intentionally or unintentionally. Regencies/cities for data collection also have different characteristics, both in terms of geographic topology and available information technology facili-

ties, in this case internet connection. An unstable internet connection allows for errors in data collection, while transportation difficulties impact the enumerator's home range.

Considering that Cooperation and MSME data meet minimum quality, both in terms of filling in values on attributes, as well as the amount of data while the potential for data errors is quite large, a proper mechanism is needed to ensure data quality is fulfilled the requirements. In order to assure the quality of the data, in this case study we propose a data quality assurance framework based on the characteristics of the data collection project and data collected as well. This framework can also be an alternative for data collection which is massive in nature, involves many parties, within a narrow time span, involves many parties and covers a wide and varied area.

2.2 Proposed Framework

2.2.1 Overview Cooperation and MSME Data Attributes

The purpose of this data collection is to produce quality and sufficient data both in terms of individual data amount and attributes value that will be used as material for formulating policies on empowering Cooperation and MSMEs. For this purpose, according to the mandate of PP No. 7 of 2021 Article 55 Paragraph (3) (Indonesia, 2021) it is required that the main data variable group in the Single Data Information System (Sistem Informasi Data Tunggal/SIDT) of Cooperation and MSMEs contains at least Business Identity and Business Actors. In its elaboration, the Ministry of Cooperation and SMEs breaks down these variables into attributes grouped into Business Actor Identity, Business/Business Entity Identity, General Business Characteristics, Human Resources, Production/Business Processes, Marketing, and Financial Status. Other attributes as additions include: product/service marketing area, type of workforce, suppliers, turnover, venture capital, and other attributes.

2.2.2 Quality Data Assurance Framework

In compiling the framework, we use two references, i.e. data requirements specification references and data quality regulatory requirements references. The reference of data requirements is Government Regulation no. 7 of 2021, whereas the reference for the quality of data fulfillment is based on five data quality rules according to Kahn (Kahn et al., 2012). The assurance data quality framework for Cooperation and SMEs is depicted as Fig. 2. The objective of complete data collection on Cooperation and SMEs is the fulfillment of the Data Quality Rule which consists of 5 rules. Data quality assurance efforts are implemented at each stage of data collection starting from preparation, implementation/execution, to final data collection. The data quality assurance function includes all necessary actions, namely prevention of possible data errors/abnormalities, anticipation of possible errors, detection of data errors and correction of erroneous data.

In each action the mechanism or tools used are in accordance with the context of the action. In the preparation stage, features/functions are implemented in the application that are used to maintain the quality of field values since they were originally entered. To minimize data errors due to human error, in addition to implementing application features/functions, during the preparation stage, intensive training was also carried out for enumerators and verifiers.

At the implementation stage, anticipatory actions of errors are carried out by implementing two stages of verification. The flow and the second phase of the verification process are carried out semi-automatically using an application. The first stage of verification is carried out by the enumerator coordinator, while the second stage of verification is carried out by verification officers at the regency/city level.

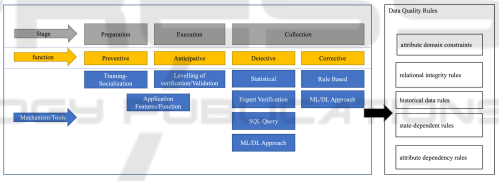

Figure 2: Proposed framework of Cooperation & MSME data quality assurance.

After the data has been collected and verified at these two stages, actions are taken to detect/identify inappropriate or abnormal data and correct errors/abnormalities. The mechanisms used at this stage are: assistance with unusual data using statistical techniques, SQL-based queries, and verification by experts in the field of microeconomics, especially Cooperation and SMEs. Detection of data that may not be normal is also carried out using ML techniques, in this case is clustering.

Data correction measures to ensure that data is correct and accurate are carried out using rule-based techniques and approaches based on ML and DL. The potential use of ML/DL is for example to assess the suitability of a place of business with the narrative of a business sector or class of business. The process of collecting and assuring data quality is presented in Figure 3.

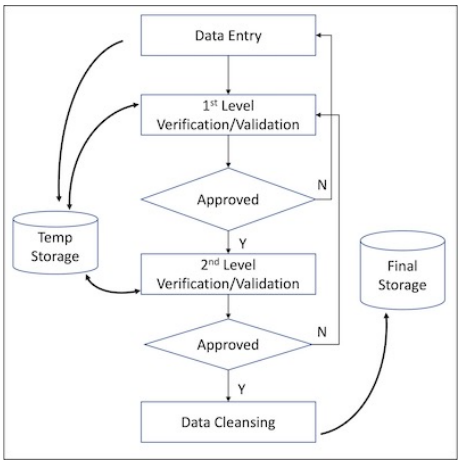

Figure 3: Flow of data quality assurance process.

3 RESULTS AND DISCUSSION

Currently, all stages of data collection implementation for this year's program have been completed. The total instant data collected is more than 9,000,0000. Assurance of data quality as in the proposed framework is already at the Data Cleansing stage, namely detective and correction. Some of the detection and correction processes have already been implemented, while other ML/DL/AI based processes are in the design stage. This section presents the results of applying some of the mechanisms and tools to the actions of the proposed framework.

3.1 Preparation

3.1.1 Application Function/ Feature

Attribute domain constraint To ensure compliance with ADC rules, we apply mandatory field features and value constraints for attributes which are the minimum requirements according to laws and regulations. Of the 237 attributes, 42 attributes are mandatory that must be filled. If these attributes are not filled, data collection cannot be continued. The mandatory attribute groups includes the business actor identity attribute, company characteristics, business location identification, business production, workforce, and production process.

The mandatory business actor identity attribute groups include the attributes of the entrepreneur's name, gender, disability (y/n), entrepreneur's status, Entrepreneur's NIK (citizen identity number), whether the address of the place of business is the same as the entrepreneur's address, province, district/city, sub-district, kelurahan/village/ nagari, complete address, telephone/cell phone number, Whats App, entrepreneur education, are you a member of a Cooperation, what type of Cooperation do you join, do you have other jobs, what other types of work do you have. The mandatory business characteristics group includes the main business/company activity attributes, the main products (goods or services) produced/sold, business entity status, initial capital at establishment, and date of operation. Mandatory business place attribute groups are province, district/city, sub-district/district, sub-district/village/nagari, full name of business/company, name of commercial/popular business, place of business, and business address. Mandatory business production attribute groups are production of goods/services produced, marketing of goods/services produced, marketing methods used, and address. The mandatory attributes of the workforce group are: wages and salaries, other incentives, number of months worked in one year, average working days per month, and average hours worked per day. Meanwhile, the production process group only attributes the use of technology in the production process.

At this stage, the guarantee of the fulfillment of the ADC is also carried out by checking by the system for attribute values that must meet certain criteria. Examples of such this attributes and their limitations are presented in Table 1.

Table 1: Attributes and their domain constraint.

Attribute	Domain Constraint
NIK (Citizen ID)	Numeric, 14 digit
No HP	Numeric, 9 – 11 digit
Kode Pos (Zip Code)	Numeric, 5 digit
Nama Pelaku Usaha (Business Actor Name)	Alphabet
Kategori Usaha (Business Category)	One Alphabet, refer to KLUI (Indonesia Business Field Category) (Keuangan, 2012)

Attribute Dependency Integrity Rule The following Table 2 presents a list of attribute dependency integrity guarantees implemented in data collection applications. The right side is an attribute that must meet the dependency rule on the left attribute. For example, the province must match the profile of the enumerator in charge of the province concerned.

3.1.2 Execution

At the data collection/data collection execution stage, the anticipation that is carried out to minimize data er-

Table 2: Attribute dependency integrity rule.

Attribute Dependency Rule
Enumerator Profile –>(Province, /City/District)
Zip Code –>Sub District
NIK (citizen identity number) –>Gender
Employee Number –>Business Status
Cooperation Member –>The type of Cooperation involved
Other business –>Profession
Finance statement –>This year income
Marketing utilizes digital media–>type of media digital used
If the business category is "G" –>(Sales income, Purchase prices of goods sold, Commission on net consignment sales)

rors/inaccuracies is to apply collection payment rules and verification. Payments to enumerators for their data collection services are based on the units of verified data they can collect. Two step of verification is implemented in stages. The first step of verification was carried out by the enumerator coordinator and the second-level verification is carried out by employees of the relevant offices in City/District Governments, as presented in Figure 2.

3.2 Collection

3.2.1 Missing Value Detection

Missing value detection is performed by using SQL query tools. As the results of database query execution, it was found approximately 6,850,000 missing value. The three attributes with the highest missing values were found in 2,519,084 of marketing method attribute, 1,606,018 of telephone number, and 1,455,625 employee. Other attributes that dominate the missing value are the business production category, business location, and business capital.

3.2.2 Expert Verification of Attribute Domain Constraint

The experts who carry out the verification at this stage are former employees of the Central Bureau of Statistics who are proficient in the concept of microeconomics. At this verification stage, it is carried out using statistical tools to identify the mean, median, minimum value and maximum value of several important attributes. Despite it has been implemented the two levels of verification in the execution phase, there are still many attributes with unrealistic values. Many cases of unrealistic values occur in business aspects such as working capital, annual business turnover, and expenses. For such those unrealistic data, the expert establish the verification rule based on the expertise, regulation, and experiences as well. Some of the verification results is presented as Table 3.

Table 3: Expert verification result of attribute domain constraint.

Block Question	Attribute Category	Business Category	Normal Value
Block 2	Working Capital (Rp.)	Individual Business	Minimum = 500.000
			Maximum = 1.000.000.000
		Non Individual (Company) Business	Minimum = 500.000.000
			Maximum = 10.000.000.000
Block 7	Omzet (Rp.)	Non Corporate	Minimum = 25.000.000
			Maximum = 200.000.000
		Corporate	Minimum = 200.000.000
			Maximum = 4.000.000.000
Block Question	Attribute Category	Business Category	Normal Value
Block 8	Number of Employee	Individual Business	Minimum = 1
			Maximum = 20
		Non Individual (Company) Business	Minimum = 20
			Maximum = 100
	Employee Salary (Rp.)	Individual Business	Minimum = 700.000
			Maximum = 3,100.000
		Non Individual (Company) Business	Minimum = 2.300.000
			Maximum = 6.300.000

3.2.3 Expert Verification of Integrity Constraints

Based on the results of the verification of experts, it was found that there was a discrepancy in the value of integrity between attributes that should be consistent. The findings of this integrity inconsistency occur in attributes related to finance and business. The main findings for this case are presented as a Table 4.

3.3 Data Correction

In the stages we have performed data correction based on rule defined and expert justification. All of data error, inconsistent, or missing value as described in sub section Missing Value Detection and Expert Verification of Attribute Domain Constraint have been corrected. Currently we are preparing to utilize the ML/DL approach to assist the error detection and corrections. Table 5 presents the initial identification of ML/DL approaches to be applied on the objective of qualified data rules.

Table 4: Attribute Dependency Integrity Rule.

Integrity Issues	# of Instant Data
Business Category is individual but its business capital >Rp. 1 B	2.725
Business Category is corporate but its business capital <500 million	6.818
Business Category is Corporate but its yearly omzet <25.000.000	3.220.942
Business Category is individual but its each employee monthly salary <Rp.	700.000 8.52.145

Table 5: ML/DL Approach Identification.

Attribute	Objective	L/DL Task (Target)	Tools	Predictor
Business Category	ADC, RIC	Classification, new- line Text Generator (Text indicates Business Category)	NN-RNN GAN-RNN	Business Image Photo, Address
Business Type	RIC	Classification (Individual, Corporate)	Classifier (Decision Tree, RNN-CNN)	Business Image Photo, Financial Aspects attributes
Working Capital	RIC	Classification (Range of WC)	Classifier (Multi Class, Decision Tree, SVM, NB)	Omzet, Number Of Employees, Business Category, Address/Location, Product, Market Segment, Target Market, product main raw materials
Type of Target Market	RIC	Classification (Type of TM)	Classifier (Multi Class, Decision Tree, SVM, NB)	Market Segment, Target Market, product main raw materials, working capial, marketing methods, Busi- ness Category, Employee (num- bers, education, salary,...)

Remark
The objective is to generate caption describes business category based on the activity/location business photo image. Many studies presents these approach are powerful to this task (Aghav, 2020; Ghandi et al., 2022; Srivastava et al., 2022) CNN based clas- sifier is the most popular technique for object-image recognition (Xuefeng Jiang, 2019; Luo et al., 2020; Lee et al., 2018; Srivastava et al., 2022; Sadikin et al., 2020; Elngar et al., 2021), where as RNN is suitable applied to a sequence data set (Olah, 2015; Sastrawan et al., 2022; Sadikin et al., 2016) For the classification task due to its predictors is quite simple, the conventional classifiers are powerful enough to overcome the "business-like" problem as pre- sented in (Sadikin et al., 2021).

4 CONCLUSIONS

The aim of the proposed framework presented in the article is to assure the quality data of Cooperation and MSMEs as the result of Complete Data Collection program run by Ministry of Cooperation and SME full the minimum standard required. The proposed framework covers all stage in the Data Collection Project by performing all activities required i.e anticipation, prevention, detection, and correction. Currently some of the stages and activities contained in the framework have been performed. As the results of the application of some parts of the framework, more than 6.8 million error data are detected and corrected.

As the Complete Data Collection Program which is continuously carried out to achieve 65 million Cooperation and MSME individual data, the assurance of data quality standard will be more complex. Therefor, we elaborate the utilization of AI approaches i.e. ML / DL. Some problems of data quality and its ML/DL approach to address the obstacle are also presented in the paper. In the next short time, we will investigate the AI approach to be applied in the program.

ACKNOWLEDGEMENTS

Thank the Deputy of Entrepreneurship Ministry of Cooperation and MSE who give us the permit the publish the lessons learned and the data cleansing case of the Program.

REFERENCES

Aghav, J. (2020). Image captioning using deep learning. *International Journal for Research in Applied Science and Engineering Technology*, 8:1430–1435.

Arndt, A., Ford, J., Babin, B., and Luong, V. (2022). Collecting samples from online services: How to use screeners to improve data quality. *International Journal of Research in Marketing*, 39:117–133.

Dziadkowiec, O., Callahan, T., Ozkaynak, M., Reeder, B., and Welton, J. (2016). Using a data quality framework to clean data extracted from the electronic health record: A case study. In *eGEMs (Generating Evidence Methods to improve patient outcomes) 4*, volume 11.

Elngar, A., Arafa, M., Fathy, A., Moustafa, B., Mahmoud, O., Shaban, M., and Fawzy, N. (2021). Image classification based on cnn: A survey. *Journal of Cybersecurity and Information Management*, 6:18–50.

Ghandi, T., Pourreza, H., and Mahyar, H. (2022). Deep learning approaches on image captioning: A review. Tech. rep. arXiv:2201.12944.

Indonesia, G. (2021). Indonesia goverment law number 07 year 2021 of ease, protection, and empowerment of cooperations and micro, small, and medium enterprises.

Kahn, M., Raebel, M., Glanz, J., Riedlinger, K., and Steiner, J. (2012). A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical Care*, 50:21–29.

Keuangan, D. P. K. (2012). Keputusan direktur jenderal pajak no. kep-321/pj/2012.

Lee, S., Chen, T., Yu, L., and Lai, C. (2018). Image classification based on the boost convolutional neural network. *IEEE Access*, 6:12755–12768.

Liu, S., Andrienko, G., Wu, Y., Cao, N., Jiang, L., Shi, C., Wang, Y., and Hong, S. (2018). Steering data quality with visual analytics: The complexity challenge. *Visual Informatics*, 2:191–197.

Luo, Y., Zhang, T., Li, P., Liu, P., Sun, P., Dong, B., and Ruan, G. (2020). Mdfi: Multi-cnn decision feature integration for diagnosis of cervical precancerous lesions. *IEEE Access*, 8:29616–29626.

Martinez-Luengo, M., Shafiee, M., and Kolios, A. (2019). Data management for structural integrity assessment of offshore wind turbine support structures: data cleansing and missing data imputation. *Ocean Engineering*, 173:867–883.

Nikiforova, A. (2020). Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. *Baltic Journal of Modern Computing*, 8:391–432.

Olah, C. (2015). Understanding lstm networks.

Ouyang, B., Song, Y., Li, Y., Sant, G., and Bauchy, M. (2021). Ebod: An ensemble-based outlier detection algorithm for noisy datasets. *KnowledgeBased Systems*, 231:107400.

Prokhorov, I. and Kolesnik, N. (2018). Development of a master data consolidation system model (on the example of the banking sector. *Procedia Computer Science*, 145:412–417.

Ridzuan, F. and Zainon, W. (2019). A review on data cleansing methods for big data. *Procedia Computer Science*, 161:731–738.

Sadikin, M., Fanany, M., and Basaruddin, T. (2016). A new data representation based on training data characteristics to extract drug name entity in medical text.

Sadikin, M., Ramayanti, D., and Indrayanto, A. (2020). The graded cnn technique to identify type of food as the preliminary stages to handle the issues of image content abundant. In *ACM International Conference Proceeding Series*, page 108–113.

Sadikin, M., SK, P., and Bagaskara, L. (2021). The application of machine learning approach to address the gpv bias on pos transaction.

Sastrawan, I., Bayupati, I., and Arsa, D. (2022). Detection of fake news using deep learning cnn–rnn based methods. *ICT Express*, 8:396–408.

Srivastava, S., Chaudhari, Y., Damania, Y., and Jadhav, P. (2022).

Xuefeng Jiang, Yikun Wang, W. L. S. L. J. L. (2019). Comparative performance evaluation for image classification. *International Journal of Machine Learning and Computing*, 9:10 18178 2019 9 6 881.

Zhang, Y. and Thorburn, P. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems*, 128:63–72.