Improving the MSMEs data quality assurance comprehensive framework with deep learning technique

Mujiono Sadikin[1], Purwanto S. Katidjan[2], Arif Rifai Dwiyanto[1], Nurfiyah[1], Ajif Yunizar Pratama Yusuf[1], Adi Trisnojuwono[3]
[1]Faculty of Computer Science, Universitas Bhayangkara Jakarta Raya, Bekasi, Indonesia
[2]Faculty of Economic and Business, Universitas Esa Unggul, Jakarta, Indonesia
[3]Deputy of Entrepreneurship, Ministry of Cooperatives and SME, Jakarta, Indonesia

Article Info

Article history:

Received Mar 28, 2024
Revised Sep 12, 2024
Accepted Sep 29, 2024

Keywords:

CNN-RNN
Data quality assurance
Deep learning
Image captioning
Small medium enterprises

ABSTRACT

In the year of 2022 the ministry of cooperatives and small and medium enterprises (SMEs) executed a complete data collection program for the cooperatives and micro small and medium enterprises (MSMEs) profile. As the complexity of the process and the uniqueness of the data characteristics, plenty of risks must be mitigated. The most challenging risk is the possibility of reduced data quality. This study is performed to validate the proposed comprehensive framework to ensure the quality data of cooperatives and MSME. The proposed framework aims to prevent, detect, repair, and recover dirty data to achieve the required data quality minimum standard. We investigated many techniques namely rule-based, selection-based, and deep learning-based. By applying the framework, 6,850,000 missing values are found and corrected, whereas the number of instant data containing attribute values that do not follow the domain constraints or integrity rule is 4,082,630. The first deep learning task applied in the framework is MSME activity image description (image captioning) generated by the convolutional neural network-recurrent neural network (CNN-RNN) model. By using 1000 MSME images as data training, the model's performance is quite good, achieving the average BLEU score of Culinary 0,3149, Fashion 0,4868, and creative products 0,5086. So far, the proposed framework can contribute to supporting MSME one data as the Indonesian government program.

Corresponding Author:

Mujiono Sadikin
Faculty of Computer Science, Universitas Bhayangkara Jakarta Raya
Bekasi, Indonesia
Email: mujiono.sadikin@gmail.com

1. INTRODUCTION

Data quality is the most important issue in decision-making in all aspects of human life. Therefore, quality data assurance has gotten more attention from many practitioners and researchers [1]. The studies and practice reports have been published to address the data quality issues, specifically in certain areas. Those publications on data quality include online sales and marketing, pharmaceutical and health [2]–[5], marine and environment monitoring [6], [7], industry manufacture [8], company registration cases in several countries [9], the internet of things (IoT) applications field [10], and concrete analysis in the infrastructure sector [11].

Numerous strategies have been proposed to guarantee that the data quality meets the requirements for generic or specific purposes. These strategies were summarized by Liu *et al.* [1]. The use of particular techniques and algorithms in data purification, which is carried out at various analytical stages depending on the data type, is also investigated by authors in their works. The authors enhance three prior general techniques by including a visualization block. Despite the proposed framework's advantage, the approach has not been tested with actual data cases. Dziadkowiec *et al.* [3] suggested a method for improving data quality that combines Kahn's work and SPSS query syntax. The authors assert that the Kahn data quality framework can enhance the resultant data quality. In this paper, we extend the option of using machine learning (ML) and deep learning (DL) techniques to the Kahn framework as part of the micro small and medium enterprises (MSME) data quality assurance framework. Another data quality framework in the software engineering field is proposed [12]. The framework supports the correct interpretation of empirical study data analysis. This framework works based on a set of data that has been collected, while we want to guarantee the quality of the data from planning, collection, and when it will be used.

In their review of big data cleansing [13], the authors summarize some methods that can be applied. Those methods are developed based on various techniques, i.e., rule-based, ML-based, and knowledge-based. However, those existing methods contain some limitations when dealing with dirty data.

Perfect-quality data in the pharmaceutical field is very crucial for human health purposes and regulatory compliance. To ensure the data integrity and end-to-end traceability of pharmaceutical manufacturing data, Leal *et al.* [2] proposed a comprehensive approach that covers end-to-end verification, data quality assurance, and intelligent data analysis. The end-to-end verification is implemented using blockchain technology, the data quality assurance is performed based on attributable legible contemporaneous original accurate (ALCOA) principles, and the intelligent data analysis is managed by MAS. The big data managed in the project is sourced from an automatic pharmaceutical manufacturing production line. Therefore, even though the data managed is heterogeneous with various formats, it is free of human intervention in data collecting. This condition is very different from the problem we must overcome in our project.

Arndt *et al.* [14] proposed another strategy for enhancing the quality of data. Their study utilizes filters to enhance the data quality of online market analysis. A crowd-sourcing service, namely Mturk, is used to provide the experiment data set. Based on the data sources, this methodology compares each of the four filter groups, i.e., direct selection, direct accuracy, statistical selection, and statistical accuracy.

Another approach suggested a user-focused, data-driven methodology [15]. The strategy consists of three parts: data objects, data quality specification, and data quality evaluation procedures. The suggested approach is validated with an open data set of companies that are registered in numerous nations, including Latvia, Norway, England, and Estonia. The authors argue that the shortcomings of the prior strategy can be fixed by using the method. Despite the proposed method being thorough, it only works with data that adheres to specific requirements, such as completeness, freedom of ambiguous values, and correctness. Due to such barriers, there are continuous problems regarding data quality that must be resolved.

The proper handling of master data is another factor that must be considered to ensure data quality [16]. Consolidation, harmonization, and management are the three components of the proposed master data model management system. The data structure and data gathering are improved during the consolidation stage. Alignment, normalization, and classification are performed during the harmonization stage, whereas centralization and administration are necessary during the management action stage.

When data collection methods are almost real-time, like in high-frequency water quality monitoring systems as studied by Zhang and Thorburn [7], it is more challenging to ensure data quality. Real-time data quality might decline due to a variety of reasons including network issues, device malfunctions, and device replacements. In their study, the authors developed a cloud-based system that integrates several methodologies and cutting-edge algorithms to execute missing value imputation in real-time data sets, overcoming the missing values that will decrease the accuracy of the information delivered. The developed system utilizes several imputation methods, such as mean imputation, last observation carried forward (LOCF), linear imputation, EM, multivariate imputations by chained equations (MICE), dual-head sequence-to-sequence imputation model (dual-SSIM), and M-RNN. When applied to nitrate and water temperature data, Dual-SSIM performs best among these methods. Even though the method effectively handles real-time data, it does not address data integrity, integrity between attributes, or interactions between attributes.

Most of the data set is not error-free, that's why the real-world data is mostly dirty. The real data set is frequently inconsistent, has missing numbers, is unreliable and unclear, and contains outliers. Since the quality of the information produced is dependent on the quality of the data, data cleaning is therefore not only the primary activity but also the most crucial component of data management [13]. Different strategies are needed to perform data cleansing tasks. Ouyang *et al.* [11], Ouyang et al. described how they employed an ML-based ensemble strategy to find outliers in a concrete measurement regression data set. In the comparison of k nearest neighbors (KNN), local outlier factor (LOF), connectivity-based outlier factor (COF), one-class support vector machines (OCSVM), IFOREST, angle-based outlier detection (ABOD), and

SOS, the authors utilize an ANN-based model approach. Based on their experiment results, the ANN-based model provides the best performance in identifying outliers in the used regression data.

The energy manufacturing sector also generates noisy data, such as the SCADA monitoring system's data on the structure health of offshore wind turbine constructions, which might lead to a mistake in decision-making. To solve the issue, [8] suggested a methodology based on artificial neural network (ANN) techniques to enhance data quality by automating data cleaning operations. The proposed approach consists of two steps: imputation for missing data and verification, and elimination of data noise. The goal of the study was to enhance the accuracy of fatigue assessments made for turbines, which are assumed to be significantly influenced by the monitoring data produced by SCADA sensors. Therefore, in the study, the authors assessed the quality of the data without and with cleaning using the suggested method. The authors are convinced that, based on the experimental findings, the data cleansing improved the quality of the decision-making.

An organizational approach to performing the data quality management (DQM) process is introduced [10]. The authors propose a framework for IoT DQM maturity consisting of two elements, i.e., a process reference model for IoT and a maturity model suitable for IoT DQM. The framework is established based on the process maturity models of ISO 8000-61 and ISO 8000-62.

The quality of the data becomes a more significant issue when the managed data is to be shared for consortia with the members spread into various locations, such as IeDEA [17]. For the data quality assurance purposes of global HIV research, the consortium implements the IeADA Harmonized Data Toolkit. The toolkit is an open-source software system that evaluates the quality of data when the data component is uploaded to the system. Quality data assurance is performed based on certain rules covering non-nullable primary keys, primary constraints, date sequences, value ranges, etc.

DL is one of the promising methods used to assure data quality, mainly in healthcare image data provided by magnetic resonance imaging (MRI) sensors, as presented by [18]–[21]. All these studies use an MRI image dataset with convolutional neural network (CNN) and its variant as the deep learning technique. Another approach to improve data quality proposed is the utilization of long short-term memory (LSTM) as published in [22]. In their work, authors apply two LSTM implementation scenarios to address the low-quality data collected by many IoT sensors. The main problem of this data is the missing value due to many reasons. The low-quality data is also found in the online panel of the research market [23]. To overcome the information bias if it is provided based on online panel data only, Ibtissam *et al.* [23] proposed the data combination and DL technique. In the data combination stage, the online panel data is combined with the social media data whereas DL used is LSTM to perform the sentiment analysis task. Regardless of the performance of all DL approaches that have been proposed, the characteristics of data-problem quality are commonly trivial. It is slightly different from the data collection problem we must address in this work, in terms of how the data is collected and how the process complexity is.

As a challenging problem, automatic image captioning applications cover various fields, including human-machine interaction [24]–[26], robotics [27], health, and medicine [28]–[30]. Ghandi *et al.* [31] published a comprehensive literature review on an image captioning study using the DL approach. According to the authors, the common DL techniques utilized in the study area are R-CNNs, RNNs, LSTMs and gated recurrent units (GRUs), and ResNet. Some other techniques utilized are GAN [32] and VSG-LSTM [24]. This study used the CNN-RNN model to generate captions and pre-trained Xception to extract the image features.

The complete data collection of cooperatives and MSMEs 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, the data collection model, which is performed manually, the various skills and knowledge of data collection officers, the complexity of data entry forms, the short time allocated, and the project management as well. Regarding area coverage, the project covers more than 240 districts in 34 provinces across Indonesia, with various topographies and land contours. The data collection is carried out manually by more than 1,000 enumerators. As with 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 exclusivity, to the author's knowledge, no model or approach can comprehensively deal with this data quality standard problem. 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 or steps specific to the case and a combination of existing or published approaches. The proposed framework also consists of various approaches, according to the case found [33].

We also validated the DL technique, namely image-captioning, to generate descriptions of MSME photo images as an alternative solution to correct missing values of MSME business activities, products, or services in the collected photo images. Our proposed DL method is based on CNN-RNN with word indexing as sentence sequence data representation. Using the BLEU score evaluation metric, the proposed method achieves an average value of 0.3990.

2. METHOD

2.1. Overview of the MSME data collection program

The data objects to be collected are the business actors of cooperatives and MSME. The ministry of cooperatives and SMEs has the responsibility and authority to provide this business group with single data, as mandated by regulations [34]. In 2022, data collection for over 9 million individual cooperatives and MSMEs by name and address has been conducted to fulfill the obligations and responsibilities of maintaining the cooperatives and MSMS single data. Micro- and small-business actors are highly dynamic and diverse in terms of business location. Therefore, the data collection in the year stage is restricted to business actors who have a permanent business location. Two hundred and thirty-seven (237) attributes grouped into 15 blocks are used to identify each data point. Those attribute data blocks are the identity of the business actor, the identity of the place of business, the identity of business characteristics, business licenses, awards received, raw materials, products or services, labor, production processes, partnerships, financial business, coaching ever received, additional notes, and information from the registrant. Each block contains a varying number of attributes, ranging from 10 to 53.

The MSME data is collected from 240 regencies, cities, and districts in 34 provinces. Each province is represented by 1 to 33 regencies, districts, or cities based on considerations of ease of transportation access and adequacy of infrastructure. The features of cities and regions in the Indonesian territory widely vary in terms of the availability of infrastructure, topography, level of education of business actors, and social economy conditions. In the 2022 year of the data collection program, districts and cities were chosen based on some considerations such as affordability, availability of data communication infrastructure, and the population of business actors.

The data collection was conducted by 1 to 1.709 enumerators assigned in each district or city. The number of enumerators varies from one to another cities based on consideration of the size of the coverage area, the estimated number of business actors in the district or city concerned, and the degree of access difficulties. The data collection process also requires coordination between agencies, ministries, and local government since the empowerment of cooperatives and MSME business actors is multi-sectors. Complete data collection is performed by utilizing web-based applications accessed via the Internet. This application is specifically prepared for the MSME single data development, including this data collection.

Due to the complexity of the complete cooperatives and MSME data collection program in terms of the enumerator's skill, the coverage areas, and the various parties involved, the possibility of data collection errors is quite large. Enumerators (data collectors) cultural and educational backgrounds are various. Despite the training and workshops regarding the process and how to use the application of complete data collection that has been performed, since different levels of expertise or skill exist, the potential for human error in the data entry stage still exists.

Each district or city is responsible for achieving a certain amount of individual data which is then distributed to enumerators. The minimum quota that must be achieved can become a trigger to causes enumerators to make mistakes, either intentionally or unintentionally. Regencies and cities have different characteristics, both in terms of geographic topology and the availability of internet connection infrastructure. Unstable internet connections increase the possibility of errors in data collection and transmission, while transportation difficulties make it difficult for enumerators to achieve the target.

2.2. Proposed framework

2.2.1. Overview cooperatives and MSME data attributes

This data collection project aims to provide data that meets the requirements in quality and quantity so that it is suitable for use as a reference for policymaking in empowering cooperatives and MSMEs. According to government regulation number 7 of 2021, article 55 paragraph (3) [34], it is required that the main data variable group in the single data information system (*sistem informasi data tunggal*/SIDT) of cooperatives and MSMEs contains at least business actors and business identities. The Ministry of cooperatives and SMEs describes these variables into attributes grouped into business actor identity, business entity identity, general business characteristics, labor, production processes, sales, and financial aspects. Other attributes that are added include products and services, marketing area, type of workforce, suppliers, turnover, venture capital, and other attributes.

2.2.2. Quality data assurance framework

The proposed framework is developed based on two references, i.e., data quality regulatory requirements and data requirements specifications. The reference to the quality of data fulfillment is based on five data quality rules according to Kahn *et al.* [35], whereas the reference to data quality regulatory requirements is government regulation No. 7 of 2021. Figure 1 depicts the assurance data quality framework of cooperatives and SMEs, whereas Figure 2 presents the process of collecting and assuring data quality.

The objective of the complete data collection on cooperatives and SMEs is to fulfill the five rules of data quality. These five rules are attribute domain constraints (ADC), relations integrity rules (RIR), historical data rules (HDR), state-dependent rule (SR), and attribute dependency rules (ADR). Data quality assurance processes are performed at each stage of data collection, starting from preparation, implementation, and execution to final data collection. The data quality assurance function includes all necessary actions, namely the prevention of possible data errors or abnormalities, the anticipation of possible errors, the detection of data errors, and the correction of erroneous data. In every stage, some mechanisms or tools are used according to the context of the action. In the preparation stage, features and functions are implemented in the application used to maintain the quality of attribute values since they were originally entered. To minimize data errors due to human mistakes, in addition to the application features and functions implementation, during the preparation stage intensive training for enumerators and verifiers was also conducted.

At the implementation stage, anticipatory actions for 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 performed by verification officers at the regency or city level.

After the data has been collected and verified at these two stages, actions are taken to detect abnormal or inappropriate data and perform the correction. The mechanisms used at this stage are statistical techniques, SQL-based queries, and verification by experts in microeconomics, especially corporations and SMEs. The detection of data that may not be normal is also carried out using clustering techniques. Data correction mechanisms to ensure data correctness and accuracy are rule-based techniques and ML and DL approaches. The rule-based data correction is performed manually by trained staff. The process of collecting and ensuring data quality is presented in Figure 2.

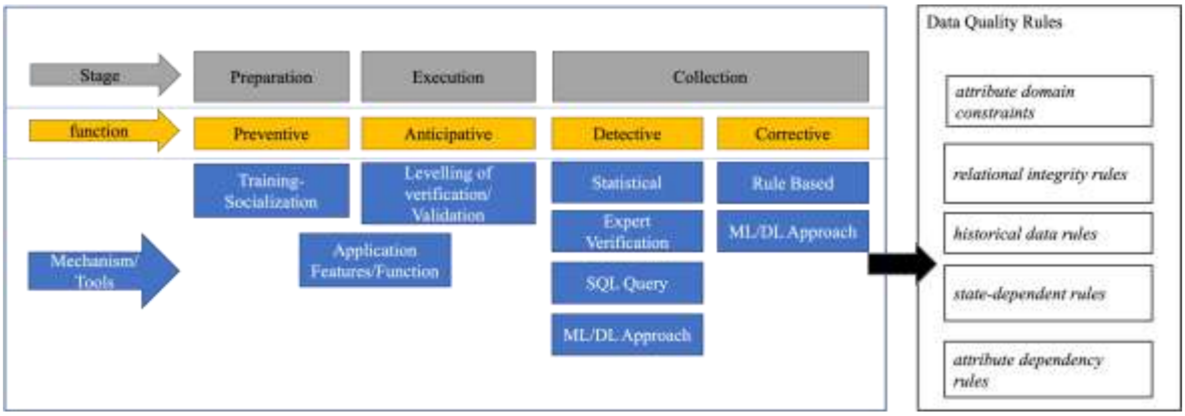

Figure 1. The proposed framework of cooperatives and MSME data quality assurance

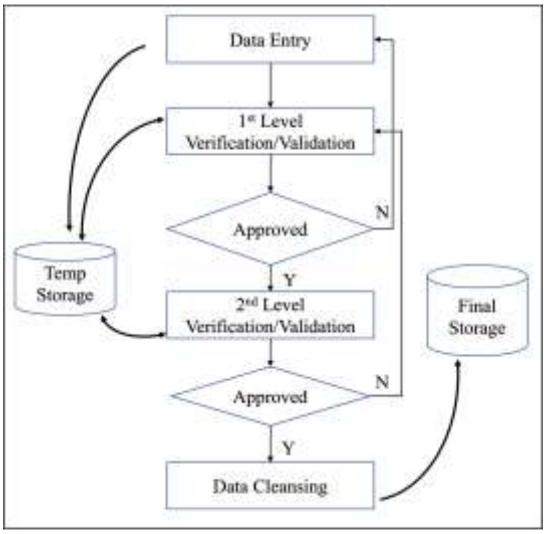

Figure 2. Flow of data quality assurance process

2.2.3. Deep learning approach and experiment

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. Currently, we have investigated some ML and DL approaches to assist in error detection and correction. Table 1 presents the initial identification of ML/DL approaches to be applied to the objective of qualified data rules.

In this experiment, we proposed a model to predict text descriptions in business activity images. This description will then be used for various identifications, including the type of business, business group, and products or services provided. Referring to Table 1, the initial experiment using DL is described in the first row. Figure 3 depicts the experiment stages used to predict the business activities' photo descriptions. Overall, the experiment contains two main blocks, i.e., data treatment and prediction model development. The data preparation part consists of data collection, data pre-processing, and data representation. In the model development stage, we perform model training and model testing.

Business activity photo images as the data set are collected from two sources, i.e., the real data provided by the enumerators that are entered into the system and those images collected by us. All those data images are then manually labeled with appropriate descriptions. We use a limited crowdsourcing approach to perform the labeling.

The data preprocessing treatment is applied to both images and text captions. The data processing applied to the image is to standardize the size to 299×299×3 as the standard input of Xception. The text preprocessing applied is to make all words in lowercase representation. The representation of text caption is the index of unique words in all image captions, thus we have vocab_size as the maximum index. Each image text caption label is then transformed into a one-dimensional vector of the word index. The vector. length is the maximum length of text captions, so we have max_length of text caption vector. If the certain text-caption length is less than max_length, it is padded with 0 value to the rest.

We use Xception as the pre-trained CNN for ImageNet [46] to extract the image features. The architecture of Xception is presented in Figure 4. The architecture consists of 3 main blocks i.e., entry flow, middle flow, and exit flow. Conventionally, the architecture is used for image classification using logistics regression as the classifier. Since we only need the image feature in the Xception model, its classification layer is removed, so we get the extracted feature as a one-dimensional 2048-length vector. The extracted image feature vector and the index of word caption representation are then used as the inputs of the proposed CNN-LSTM model depicted in Figure 5. The model architecture mainly consists of 3 blocks i.e., the CNN, LSTM, and decoder-output layer. CNN is used to handle image features, LSTM is used to hold text caption sequences, and the decoder combines CNN-LSTM. The model output is stored in a one-dimensional vector of word indices of length vocab_size.

Table 1. ML/DL approach identification

Attribute	Objective (data quality rule)	Ml/Dl task (target)	Tools	Predictor	Remark
Business category	ADC, RIR,	Classification, text generator (text indicates business category)	CNN-RNN GAN-RNN	Business image photo, address	The objective is to generate caption describes business category based on the activity / location business photo image. Many studies present this approach are powerfull to this task [36], [31]
Business type	RIR	Classification (individual, corporate)	Classifier (decision tree, RNN-CNN)	Business image photo, financial aspects attributes	CNN based classifier is the most popular technique for obect-image recognition [37]–[41], whereas RNN is suitable applied to a sequence dataset [42]–[44]
Working capital (WC)	RIR	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,	For the classification task due to its predictors is quite simple, the conventional classifiers are powerfull enough to overcome the "busines-like" problem as presented in [45]
Type of target market (TM)	RIR	Classification (type of TM)	Classifier (multi class, decision tree, SVM, NB)	Market segment, target market, product main raw materials, working capial, marketing methods, business category, employee (numbers, education, salary)	

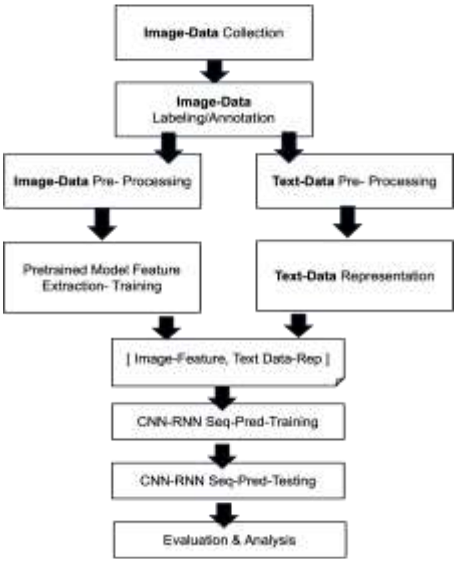
Image-Data Collection
Image-Data Labeling/Annotation
Image-Data Pre- Processing
Text-Data Pre- Processing
Pretrained Model Feature Extraction- Training
Text-Data Representation
[Image-Feature, Text Data-Rep]
CNN-RNN Seq-Pred-Training
CNN-RNN Seq-Pred-Testing
Evaluation & Analysis

Figure 3. Experiment stages of the DL approach to generate MSME photo activity descriptions

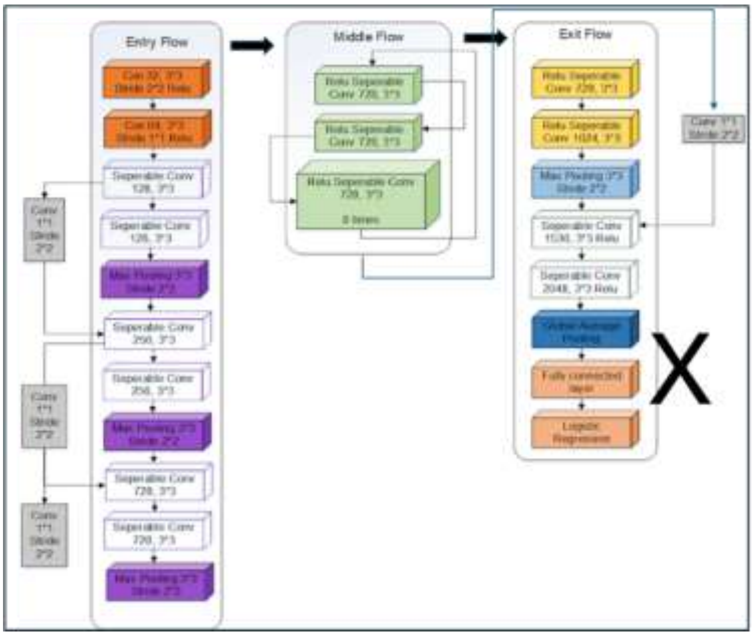
Entry Flow
Conv 32, 3*3 Stride 2*2 Relu
Conv 64, 3*3 Stride 1*1 Relu
Seperable Conv 128, 3*3
Seperable Conv 128, 3*3
Max Pooling 3*3 Stride 2*2
Seperable Conv 256, 3*3
Seperable Conv 256, 3*3
Max Pooling 3*3 Stride 2*2
Seperable Conv 728, 3*3
Seperable Conv 728, 3*3
Max Pooling 3*3 Stride 2*2
Conv 1*1 Stride 2*2
Conv 1*1 Stride 2*2
Conv 1*1 Stride 2*2
Middle Flow
Relu Seperable Conv 728, 3*3
Relu Seperable Conv 728, 3*3
Relu Seperable Conv 728, 3*3
8 times
Exit Flow
Relu Seperable Conv 728, 3*3
Relu Seperable Conv 1024, 3*3
Max Pooling 3*3 Stride 2*2
Seperable Conv 1536, 3*3 Relu
Seperable Conv 2048, 3*3 Relu
Global Average Pooling
Fully connected layer
Logistic Regression
Conv 1*1 Stride 2*2
X

Figure 4. Xception CNN architecture for feature extraction and classification

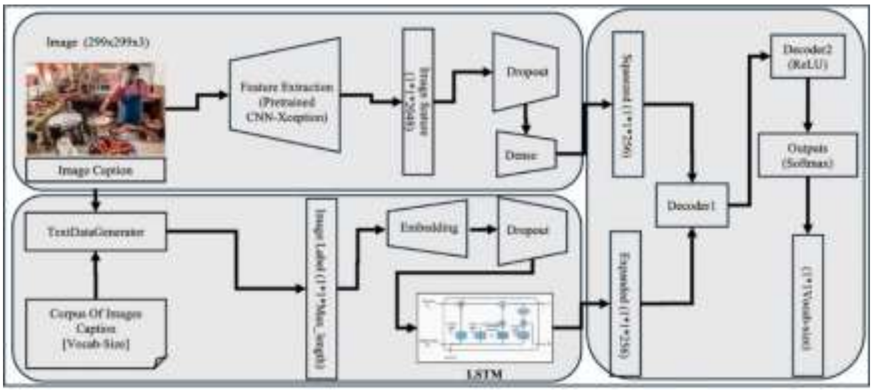

Figure 5. The CNN-RNN model proposed

To evaluate the model performance in generating the MSME activity image description, we use the BLEU parameter as computed in the below formula [47].

$$P = \frac{\sum_{n-gram \in (y,y')} Count_{clip}(n-gram)}{\sum_{n-gram \in y'} Count(n-gram)} \quad (1)$$

Where P is the BLEU precision score, $Count_{clip}$ represents the maximum number of appearing n-grams in both the reference and predicted description as well, whereas Count denotes the number of n-grams appearing in the predicted description. Since the BLEU is only able to evaluate the syntax caption, it is needed to manually analyze the semantic part of the description.

3. RESULTS AND DISCUSSION

All stages of data collection implementation have been completed. The total instant data collected is more than 9,000,000. In the data cleaning stage, namely detective and correction some of the detection and correction processes have already been implemented, whereas the initial ML/DL/AI-based experiment processes are described in the MSME activity image description as presented in the paper. 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 for ADC

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 are mandatory and must be filled. If these attributes are not filled, data collection cannot be continued. The mandatory attribute groups include 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, WhatsApp, entrepreneur education, Are you a member of certain cooperatives? What type of cooperatives 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 or company activity attributes, the main products (goods or services) produced or sold, business entity status, initial capital at the 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 and services produced, marketing of goods and services produced, marketing methods used, and address. The mandatory attributes of the workforce group are wages and salaries, other incentives, the number of months worked in a year, average working days per month, and average hours worked per day. Meanwhile, the production process group only attributes the use of technology to the production process.

At this stage, the guarantee of the fulfillment of the ADC is also carried out by checking the system for attribute values that must meet certain criteria. Examples of such attributes and their limitations are presented in Table 2.

Table 2. Attributes and their domain constraint

Attribut	Domain constraint
NIK (citizen ID)	Numeric, 14 digits
No HP	Numeric, 9 – 11 digits
Kode pos (Zip Code)	Numeric, 5 digits
Nama pelaku usaha (business actor name)	Alphabet
Kategori usaha (business category)	One alphabet, refer to KLUI (indonesia business field category) [48]

3.1.2. Application function/feature for ADR

Table 3 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.

Table 3. Attribute dependency integrity rule

<Independend attributes>→<dependent attributes>
Enumerator profile → (Province, /Kab/Kota/City)
Zip code → Sub district
NIK →gender
Employee number → business status
Cooperatives member → the type of cooperatives 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)

3.2. Execution

At the data collection/data collection execution stage, the anticipation that is carried out to minimize data errors and 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 steps of verification are implemented in stages. The first step of verification was carried out by the enumerator coordinator, and the second-level verification was carried out by employees of the relevant offices in city or district governments, as presented in Figure 3.

3.3. Collection

3.3.1. Missing values detection

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

3.3.2. 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 the implementation of 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 unrealistic data, the expert establishes the verification rule based on their expertise, regulation, and experiences as well. Some of the verification results are presented in Table 4.

Table 4. Expert verification results of attribute domain constraint

Block question	Attribute category	Domain value constraint	
		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 8	Number of employees	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

3.3.3. Expert verification of integrity constraints

Based on the results of the verification by 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 in Table 5.

Table 5. 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	852,145
Total	4,082,630

3.3.4. Data correction: ML/DL approach experiment

A. Data collection and annotation

MSME business categories cover a wide range of areas, and the categorization is various as well. One of the MSME categorizations is grouping the MSME businesses into 12 kinds based on the products and services they deliver. These 12 kinds of MSME businesses are culinary, fashion, agriculture, education, automotive, tours and travel, creative products, internet, health and beauty, cleaning services, kid apparel, and grocery/retail. In the initial ML/DL experiment, we use the first three categories to train and validate the CNN-RNN model. Each category contains 300 instant image data points. Since the subject highlighted is MSME activities, all images must contain the human-activity object and the produced product or service. All images are then labeled with a certain description in Bahasa that represents the activity contained in the image. Some of the images and their description labels are presented in Table 6.

Table 6. Samples of images and their descriptions label

Image	Description
	Orang menjual makanan di warung tegal (Someone is selling food in warung tegal) *Orang menjual sayur, gorengan, minuman botol di warung tegal* (Someone is selling vegetable, fried snack, drinking in the bottle in warung tegal) *Seorang wanita menjual sayur terong, ayam goreng, sop di warung tegal* (A woman is selling vegetable, fried chicken, soup in the bottle in warung tegal)
	Karyawan UMKM memasak produk untuk dijual (An employee is cooking certain product to be sold) *Karyawan UMKM menjual pisang goreng di ruko* (An employee is selling fried bananas in a store)
	seorang laki-laki sedang jual bakso di pinggir jalan (A man is selling bakso/meat ball in the street) *seorang pedagang bakso sedang keliling jualan di siang hari* (A man is selling bakso the afternoon)
	Dua orang perempuan sedang berjualan pastel di pinggir jalan (Two women are selling pastel in the street) *Dua orang ibu sedang jualan pastel di sore hari* (Two moms are selling pastel in the afternoon)

B. Data collection and annotation

Of 12 MSME categories, we use three categories, i.e., culinary, fashion, and creative products. Each used category contains 300 images as the dataset. To evaluate the proposed CNN-RNN model, we use the random split (RS) technique to split the data into data training and data testing. The composition of each category dataset for data training and data testing is 90:10, 80:20, and 70:30. The BLEU score of testing data provided by the proposed approach is presented in Table 7.

Table 7. The BLEU score performance

Category	Index	RS 90:10	RS 80:20	RS 70:30
Culinary	mean	0.2711	0.3149	0.3137
	std	0.1221	0.1400	0.1451
	min	0.0356	0.0711	0.0344
	max	0.5372	0.6069	0.7380
Fashion	mean	0.4649	0.3123	0.4868
	std	0.1335	0.0840	0.1635
	min	0.2246	0.1203	0.0896
	max	0.7236	0.5373	0.8516
Creative products	mean	0.4329	0.4861	0.5086
	std	0.1269	0.1588	0.1227
	min	0.1448	0.1448	0.1448
	max	0.7858	0.8574	0.8201

The BLEU score varies from one category to another. Some reasons we have investigated are the variation in the number of words in training data for each category and the quality of the descriptions prepared by descriptors. Due to the study using a word-indexing technique to represent the sequence of description sentences, the model is not able to predict words that are not contained in the training data. The ability of descriptors to understand image context and description writing is also different from one to another. The condition causes the quality of the image description in each category to vary and eventually causes the BLEU scores achieved to be quite different.

The BLEU performance metric measures the suitability of the number of n-grams between the original and the predicted description only. To the best of the author's knowledge, it has not been possible to evaluate the suitability of the syntax, semantics, or context. In this study, syntactic and context suitability analyses were carried out manually. Analysis was carried out on several samples of prediction results with a BLEU score around minimum, mean, and maximum values. The predicted description in general can be categorized into three groups based on the image context. The first group is the prediction of syntactic errors. The predicted description has no meaning since, structurally, the sentence is wrong. This kind of predicted description generally provides a low BLEU score, around the minimum value. The second group is the predicted descriptions that present a correct sentence structure but do not present the image context correctly. The second group in general provides the BLEU score a few points above the mean value. The third group contains predicted descriptions that are correct in both syntax and context. The sentences represented by the predicted description are generally not the same as the reference description, but the context matches the image. To the best of our knowledge, there have been no studies on image-captioning for practical purposes, especially in the MSME domain. Most studies on this use open datasets. Despite the differences in the datasets used, we present a comparison of the results of this study with the state of the art, as shown in Table 8. Compared with the state of the art, the model we propose is not too bad, although not the best. Potential improvements to our model can be made by improving the quality of the reference text description, using other data representation techniques such as word-vec/glove, or using pre-trained models other than Xception.

Table 8. BLEU score comparison of the proposed model and the state of the art

Ref.	Methods	Dataset	Avg. BLEU score	Remark
[24]	VSG-LSTM	V-CCO	0.3300	
[25]	TSG-GCN - LSTM	MSCOCO	0.3450	
[26]	TSG (topic scene graph) Faster R-CNN-LSTM	Visual genome, MSCOCO	0.8400	Max value
[29]	Contextual keyword encoder-LSTM	DeepEyeNet (DEN)	0.2030	
[32]	Generator pretraining-GAN, word-idx	SentiCap	0.6170	
[36]	InceptionV3, RNN	Flickr30K	0.6100	
	Proposed method (pretrained xception, CNN-RNN, word-idx)	MSME	0.3990	

3.3.5. Limitation

Despite its contribution to overcoming the real-world problem of data quality assurance, the study contains some limitations. The data error detection is still manually performed by a query, not in UI/UX style. Since time limitations constrain this stage, the priority is ensuring data correctness. At the same time, user convenience is the second priority more and less for the same reason presented by Lewis *et al.* [17], which also perform error detection manually. Another limitation is that the DL model proposed is unable to predict the words not contained in the training description vocabulary. The limitation is embedded in the word-indexing weakness, which can consider only all words in the training data to represent the text sequence in the numeric format.

4. CONCLUSION

Data quality that meets standards is an absolute requirement to provide correct information for decision-making. The proposed framework presented in the article aims to ensure the quality of data for cooperatives and MSMEs as the basis of Indonesian government policymaking regarding the MSMEs. The proposed framework covers all stages of the data collection project by performing all activities required, i.e., anticipation, prevention, detection, and correction. Some of the stages and activities in the framework have been performed. As a result of the application of some parts of the framework, more than 6.8 million error data points were detected and corrected. Without the proper framework, it is impossible to assure the quality of millions of instant data that are manually collected. As the complete data collection program is continuously carried out to achieve 65 million cooperatives and MSME individual data, the assurance of data quality standards will become more complex. Therefore, we elaborate on the utilization of AI approaches, i.e., ML and DL. Some problems with data quality and its ML/DL approach to address the obstacles are also presented in the paper. The initial stage of the application of the DL approach presented in the study is the CNN-RNN technique to generate an MSME business activity image description. Regardless of its limitations, the experiment with the DL approach convinced us that this technique is feasible for performing the data imputation on the missing value of the image description. To improve the framework performance, especially in the DL technique we are investigating the other text representation technique namely word embedding. The advantage of this technique is the possibility to utilize external knowledge, i.e., others' corpus, to improve the quality of text features. The other possibility is to enrich the dataset as we have more than 8 million photo images of MSME business activities. In the DL implementation, commonly the more instant data quantity used will provide a better model.

ACKNOWLEDGEMENTS

The authors thank the deputy of entrepreneurship, ministry of cooperatives and SME, who gave us the permit to publish the lessons learned and the data cleansing case of the program.

REFERENCES

[1] S. Liu *et al.*, "Steering data quality with visual analytics: the complexity challenge," *Visual Informatics*, vol. 2, no. 4, pp. 191–197, Dec. 2018, doi: 10.1016/j.visinf.2018.12.001.

[2] F. Leal *et al.*, "Smart pharmaceutical manufacturing: ensuring end-to-end traceability and data integrity in medicine production," *Big Data Research*, vol. 24, p. 100172, May 2021, doi: 10.1016/j.bdr.2020.100172.

[3] O. Dziadkowiec, T. Callahan, M. Ozkaynak, B. Reeder, and J. Welton, "Using a data quality framework to clean data extracted from the electronic health record: a case study.," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*, vol. 4, no. 1, p. 11, Jun. 2016, doi: 10.13063/2327-9214.1201.

[4] A. Daneshkohan, M. Alimoradi, M. Ahmadi, and J. Alipour, "Data quality and data use in primary health care: a case study from Iran," *Informatics in Medicine Unlocked*, vol. 28, p. 100855, 2022, doi: 10.1016/j.imu.2022.100855.

[5] A. Sohrabi, A. Sabahi, A. Garavand, and L. Ahmadian, "Data validation techniques used in admission discharge and transfer systems: Necessity of use and effect on data quality," *Informatics in Medicine Unlocked*, vol. 34, p. 101122, 2022, doi: 10.1016/j.imu.2022.101122.

[6] J. Xie, H. Jiang, W. Song, and J. Yang, "A novel quality control method of time-series ocean wave observation data combining deep-learning prediction and statistical analysis," *Journal of Sea Research*, vol. 195, p. 102439, Oct. 2023, doi: 10.1016/j.seares.2023.102439.

[7] Y. Zhang and P. J. Thorburn, "Handling missing data in near real-time environmental monitoring: a system and a review of selected methods," *Future Generation Computer Systems*, vol. 128, pp. 63–72, Mar. 2022, doi: 10.1016/j.future.2021.09.033.

[8] M. Martinez-Luengo, M. Shafiee, and A. Kolios, "Data management for structural integrity assessment of offshore wind turbine support structures: data cleansing and missing data imputation," *Ocean Engineering*, vol. 173, pp. 867–883, Feb. 2019, doi: 10.1016/j.oceaneng.2019.01.003.

[9] A. Nikiforova, "Definition and evaluation of data quality: user-oriented data object-driven approach to data quality assessment," *Baltic Journal of Modern Computing*, vol. 8, no. 3, pp. 391–432, Sep. 2020, doi: 10.22364/BJMC.2020.8.3.02.

[10] S. Kim, R. Pérez-Castillo, I. Caballero, and D. Lee, "Organizational process maturity model for IoT data quality management," *Journal of Industrial Information Integration*, vol. 26, p. 100256, Mar. 2022, doi: 10.1016/j.jii.2021.100256.

[11] B. Ouyang, Y. Song, Y. Li, G. Sant, and M. Bauchy, "EBOD: an ensemble-based outlier detection algorithm for noisy datasets," *Knowledge-Based Systems*, vol. 231, p. 107400, Nov. 2021, doi: 10.1016/j.knosys.2021.107400.

[12] M. M. Rosli and N. S. M. Yusop, "Evaluating the effectiveness of data quality framework in software engineering," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 6, pp. 6410–6422, Dec. 2022, doi: 10.11591/ijece.v12i6.pp6410-6422.

[13] F. Ridzuan and W. M. N. Wan Zainon, "A review on data cleansing methods for big data," *Procedia Computer Science*, vol. 161, pp. 731–738, 2019, doi: 10.1016/j.procs.2019.11.177.

[14] A. D. Arndt, J. B. Ford, B. J. Babin, and V. Luong, "Collecting samples from online services: How to use screeners to improve data quality," *International Journal of Research in Marketing*, vol. 39, no. 1, pp. 117–133, Mar. 2022, doi: 10.1016/j.ijresmar.2021.05.001.

[15] A. Fernández and F. Llorens, "An IT governance framework for universities in Spain," *EUNIS 2009 Conference*, pp. 1–13, 2009.

[16] I. Prokhorov and N. Kolesnik, "Development of a master data consolidation system model (on the example of the banking sector)," *Procedia Computer Science*, vol. 145, pp. 412–417, 2018, doi: 10.1016/j.procs.2018.11.093.

[17] J. T. Lewis *et al.*, "The IeDEA harmonist data toolkit: A data quality and data sharing solution for a global HIV research consortium," *Journal of Biomedical Informatics*, vol. 131, p. 104110, Jul. 2022, doi: 10.1016/j.jbi.2022.104110.

[18] D. Zukic *et al.*, "Medical image quality assurance using deep learning," *MIDL 2022 Conference*, 2022, [Online]. Available: https://github.com/OpenImaging/miqa,

[19] O. Esteban, D. Birman, M. Schaer, O. O. Koyejo, R. A. Poldrack, and K. J. Gorgolewski, "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites," *PLoS ONE*, vol. 12, no. 9, p. e0184661, Sep. 2017, doi: 10.1371/journal.pone.0184661.

[20] R. A. Pizarro *et al.*, "Automated quality assessment of structural magnetic resonance brain images based on a supervised machine learning algorithm," *Frontiers in Neuroinformatics*, vol. 10, Dec. 2016, doi: 10.3389/fninf.2016.00052.

[21] A. D. Pontoriero *et al.*, "Automated data quality control in FDOPA brain PET imaging using deep learning," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106239, Sep. 2021, doi: 10.1016/j.cmpb.2021.106239.

[22] C. Hwang, K. Lee, and H. Jung, "Improving data quality using a deep learning network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 1, pp. 306–312, Oct. 2020, doi: 10.11591/ijeecs.v20.i1.pp306-312.

[23] Y. Ibtissam, A. Abdallah, and H. Mohamed, "Online panel data quality: a sentiment analysis based on a deep learning approach," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 3, pp. 1468–1475, 2023, doi: 10.11591/ijai.v12.i3.pp1468-1475.

[24] K. Nguyen, S. Tripathi, B. Du, T. Guha, and T. Q. Nguyen, “In defense of scene graphs for image captioning,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 1387–1396. doi: 10.1109/ICCV48922.2021.00144.

[25] D. Wang, D. Beck, and T. Cohn, “On the role of scene graphs in image captioning,” in *LANTERN@EMNLP-IJCNLP 2019 - Beyond Vision and LANguage: inTEgrating Real-World kNowledge, Proceedings*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 29–34. doi: 10.18653/v1/d19-6405.

[26] M. Zhang, J. Chen, P. Li, M. Jiang, and Z. Zhou, “Topic scene graphs for image captioning,” *IET Computer Vision*, vol. 16, no. 4, pp. 364–375, Jun. 2022, doi: 10.1049/cvi2.12093.

[27] R. C. Luo, Y. T. Hsu, Y. C. Wen, and H. J. Ye, “Visual image caption generation for service robotics and industrial applications,” *Proceedings - 2019 IEEE International Conference on Industrial Cyber Physical Systems, ICPS 2019*, pp. 827–832, 2019, doi: 10.1109/ICPHYS.2019.8780171.

[28] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, “A survey on biomedical image captioning,” in *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 26–36. doi: 10.18653/v1/W19-1803.

[29] J. H. Huang, T. W. Wu, and M. Worring, “Contextualized keyword representations for multi-modal retinal image captioning,” in *ICMR 2021 - Proceedings of the 2021 International Conference on Multimedia Retrieval*, New York, NY, USA: ACM, Aug. 2021, pp. 645–652. doi: 10.1145/3460426.3463667.

[30] D. R. Beddiar, M. Oussalah, and T. Seppänen, “Automatic captioning for medical imaging (MIC): a rapid review of literature,” *Artificial Intelligence Review*, vol. 56, no. 5, pp. 4019–4076, May 2023, doi: 10.1007/s10462-022-10270-w.

[31] T. Ghandi, H. Pourreza, and H. Mahyar, “Deep learning approaches on image captioning: a review,” ACM comput. surv., vol. 56, no. 3, 2022, [Online]. Available: http://arxiv.org/abs/2201.12944

[32] D. Setiawan, M. A. C. Saffachrissa, S. Tamara, and D. Suhartono, “Image Captioning with Style Using Generative Adversarial Networks,” *International Journal on Informatics Visualization*, vol. 6, no. 1, pp. 26–32, 2022, doi: 10.30630/joiv.6.1.709.

[33] M. Sadikin, A. Trisnojuwono, . Rudiansyah, and T. Widodo, “Comprehensive Approach to Assure the Quality of MSME Data in Indonesia: A Framework Proposal,” pp. 154–161, 2024, doi: 10.5220/0012445700003848.

[34] G. Indonesia, Indonesia Government Law Number 07 Year 2021 Of Ease, Protection, and Empowerment of Cooperatives and Micro, Small, and Medium Enterprises, no. 086507. Indonesia: https://jdih.setkab.go.id/PUUdoc/176384/PP_Nomor_7_Tahun_2021.pdf, 2021, pp. 1–121. [Online]. Available: https://jdih.setkab.go.id/PUUdoc/176384/PP_Nomor_7_Tahun_2021.pdf

[35] M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner, “A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research,” *Medical Care*, vol. 50, no. SUPPL. 1, pp. S21–S29, Jul. 2012, doi: 10.1097/MLR.0b013e318257dd67.

[36] J. Aghav, “Image Captioning using Deep Learning,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, no. 6, pp. 1430–1435, Jun. 2020, doi: 10.22214/ijraset.2020.6232.

[37] X. Jiang, Y. Wang, W. Liu, S. Li, and J. Liu, “CapsNet, CNN, FCN: Comparative performance evaluation for image classification,” *International Journal of Machine Learning and Computing*, vol. 9, no. 6, 2019, doi: 10.18178/ijmlc.2019.9.6.881.

[38] Y. M. Luo *et al.*, “MDFI: multi-CNN decision feature integration for diagnosis of cervical precancerous lesions,” *IEEE Access*, vol. 8, pp. 29616–29626, 2020, doi: 10.1109/ACCESS.2020.2972610.

[39] S. J. Lee, T. Chen, L. Yu, and C. H. Lai, “Image classification based on the boost convolutional neural network,” *IEEE Access*, vol. 6, pp. 12755–12768, 2018, doi: 10.1109/ACCESS.2018.2796722.

[40] M. Sadikin, D. Ramayanti, and A. P. Indrayanto, “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*, New York, NY, USA: ACM, Feb. 2020, pp. 108–113. doi: 10.1145/3384613.3384649.

[41] A. Elngar *et al.*, “Image classification based on CNN: a survey,” *Journal of Cybersecurity and Information Management*, pp. 18-50, 2021, doi: 10.54216/jcim.060102.

[42] C. Olah, “understanding LSTM networks,” [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[43] I. K. Sastrawan, I. P. A. Bayupati, and D. M. S. Arsa, “Detection of fake news using deep learning CNN–RNN based methods,” *ICT Express*, vol. 8, no. 3, pp. 396–408, Sep. 2022, doi: 10.1016/j.icte.2021.10.003.

[44] M. Sadikin, M. I. Fanany, and T. Basaruddin, “A new data representation based on training data characteristics to extract drug name entity in medical text,” *Computational Intelligence and Neuroscience*, vol. 2016, pp. 1–16, 2016, doi: 10.1155/2016/3483528.

[45] M. Sadikin, S. K. Purwanto, and L. R. Bagaskara, “The application of machine learning approach to address the GPV bias on POS transaction,” *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 14, pp. 3428–3438, 2021.

[46] F. Chollet, “Xception: Deep Learning with Depth wise Separable Convolutions,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.

[47] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, vol. 2002-July, pp. 311–318, 2002.

[48] M. of F. Directorat General of Taxes, Decree of Taxes Directorate General, No. Kep-321/PJ/2012. 2012, pp. 1–94. [Online]. Available: https://www.pajakku.com/tax-guide/9855/KEP_DIRJEN_PJK/KEP-321/PJ/2012

BIOGRAPHIES OF AUTHORS

Mujiono Sadikin is the associate professor at Universitas Bhayangkara Jakarta Ratya. He acquired his bachelor’s degree in informatics in 1995 and Magister’s degree in Informatics in 2000, from Bandung Institute of Technology, followed by a doctoral degree in Computer Science from Universitas Indonesia, Jakarta in 2017. He got several research grants from the Indonesian government from 2012 to 2020. He published more than articles and several books. He holds several professional certifications i.e. CISA, CGEIT of ISACA, and ASEAN Engineer. His area of research includes AI, Machine Learning, and IT Governance. The focus of his publications is mainly machine learning and data science. He can be contacted at email: mujiono.sadikin@gmail.com.

Purwanto S. Katidjan undergraduate education was taken at the Agribusiness Study Program, Bogor Agricultural Institute—IPB (1992). Master's education continued at the Agricultural Economics Study Program, IPB (2000). Doctoral education was taken at the Business Management Doctoral Program, IPB (2011). Currently, he is member of the Economic Business Faculty of Universitas Esa Unggul Jakarta. He published many articles and textbooks in statistics and economic areas. His research interests in entrepreneurship, human resources, and statistics. He can be contacted at email: purwanto_sk@ueu.ac.id.

Arif Rifai Dwiyanto He acquired his bachelor's degree in informatics in 1999 from Bandung Institute of Technology and Magister's degree in Information Technology from Universitas Indonesia in 2016. Now he is seeking opportunities to continue his education to a doctoral degree. He is a Faculty of Computer Science Universitas Bhayangkara Jakarta Raya member. He holds several professional certifications mainly in IT development, cloud computing, and IS audit. His area of research includes cyber security, AI, Machine Learning, IT Infrastructure, and IT Governance. The focus of his publications is mainly security, machine learning, and data science. He can be contacted at email: arif@ubharajaya.ac.id.

Nurfiyah she obtained a Bachelor of Science degree in Informatics in 2016 at Bhayangkara University of Jakarta Raya and a Master of Science in Computer Science in 2020 from STMIK Nusa Mandiri. She received a research grant from the Indonesian government in 2021. She published some articles. He holds several professional certifications namely Database Administrator. Her research fields include Data Science, Image Processing, and IT Governance. Her publications focus mainly on data science and image processing. She can be contacted at email: nurfiyah@dsn.ubharajaya.ac.id.

Ajif Yunizar Pratama Yusuf is currently a lecturer in the faculty of computer science at Universitas Bhayangkara Jakarta Raya. He received his bachelor's degree in mathematics from Hasanuddin University in 2011. In 2018, he received his master's degree in computer science from Kyushu Institute of Technology, Japan. His research area includes artificial intelligence, image processing, and machine learning. He can be contacted at email: ajif.yunizar@dsn.ubharajaya.ac.id.

Adi Trisnojuwono He acquired his bachelor's degree in Agriculure Social Economif from Universitas Halu Oleo, Kendari 1992, and his master's degree is acquired master's degree in SME Professional Industry from Institut Pertanian Bogor, 2017. Currently, he acts as the Deputy Assistance of Entrepreneurship Deputy, the Ministery of Cooperatives and SME. He can be contacted via email: aditrisnojuwono@gmail.com.