International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Ensuring Data Reliability in AI Systems: Connecting Data Quality and Model Integrity

Sunil Kumar Mudusu
AI Data Engineer | Highmark Health Solutions | Pittsburgh, PA, USA

Abstract - The paper is based on an AI-based system analysis that helps to ensure the reliability with the use of quality of data and models. The data reliability engineering is different type of system compared to traditional data quality management. In this process reliability contracts, anomaly detection SLAs, and lineage-aware monitoring are play a vital role. Various case study like property buying data has been used in this research work to understand the quality like how much missing data and anomalies are present and how this can effect to the analysis process. Various machine learning farework such as Linear Regression, Decision Trees and Random Forest are used with their accuracy, Precision, and F1-score to evaluate these models. The ethical issues and main limitations af the study are also analysis in this research. The results can help to enhance the stability and transparency of AI system with sustainable and valid model performance.

Keywords: data reliability, AI systems, data quality, model integrity, anomaly detection, lineage

I. INTRODUCTION

AI systems are becoming part of businesses, and the accuracy of upstream information is paramount. Data reliability engineering (DRE) is critical in improving the level of trust in the model and reducing model degradation. Conventional data quality measures are ineffective in tackling the real-time issues that AI systems face. In this paper, we present DRE, which is based on reliability contracts, anomaly detection, and lineage-aware monitoring. Case studies illustrate the improvement of model integrity and performance by DRE practices in real-world environments.

Research Aim and Objectives

Aim

The paper focuses on making Data Reliability Engineering (DRE) an important field in ensuring the data integrity of AI systems.

Objectives

- To establish the meaning of Data Reliability Engineering and what differentiates it from conventional data quality management practices.
- To implement the reliability contract, anomaly detection SLAs and lineage-aware monitoring into the DRE practices.
- To assess how the DRE practices can affect the model integrity and minimise silent model degradation.
- To introduce the case studies that show how DRE can be applied in a real-world environment of AI production.

Problem statement

Accuracy in AI systems critically depends on quality data, but the conventional data quality methods are not sufficient. Problems with real-time data reliability may result in model degradation, which is not detected, affecting the performance of AI systems [1]. These issues are commonly not observed, which results in weakened results and undermines the faith in AI models. These data reliability gaps are essential to bridge to deliver consistent, trustworthy, and reliable AI-based decisions in industries.

Novel contribution

This paper presents a new field of study called Data Reliability Engineering (DRE) that is concerned with model integrity. DRE offers a different approach to data quality in comparison to conventional data quality practices, focusing on data reliability, proactive monitoring and maintenance [2]. Some of the critical practices are reliability contracts, anomaly detection SLAs, and lineage-aware monitoring. This paper will show how DRE practices can avoid silent model degradation, which will result in more reliable and transparent AI systems by giving real-world case studies.

II. LITERATURE REVIEW

The difference between Traditional Data Quality Management and Data Reliability Engineering

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

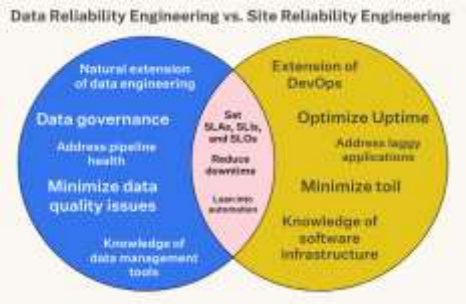

Fig 1: Comparison Data Reliability Engineering and Site Reliability Engineering

The main steps of traditional Data Quality management are precision, completeness, and congruency of the data that has been employed in AI models. DQM is used to ensure that the used data is valid and reliable upon collection [3]. Nevertheless, it tends to neglect continuous data integrity when processing AI systems in real-time. Conversely, Data Reliability Engineering (DRE) builds upon DQM by considering how to maintain reliability of data across its AI model lifecycle [4]. DRE combines proactive monitoring practices, anomaly detection, and feedback loops to identify issues before they can impact performance of the model. In contrast to DQM that is more oriented towards the properties of the static data, DRE is concerned with the dynamic problems that come up due to the constant flow of data and environment changes [5]. By refocusing the emphasis on data quality to data reliability over time, DRE will mitigate silent degradation in AI models that DQM does not handle.

AI Model Integrity Ensuring Consistent and Reliable Performance in High-Stakes Applications

Model integrity guarantees that an AI model will do what it is expected to do with consistency and reliability in its results. It has connections to the capability of the model to generalise training data to unseen new data. In case of a breach in the quality of the data, the integrity of the model is also threatened later resulting in overfitting, bias, and inefficient performance in practical scenarios [6].

Local predictions are likely to be incorrect so when model integrity is broken, an AI system may make inaccurate predictions, potentially with overwhelmingly serious consequences, particularly in high stakes areas, such as healthcare, finance, or law enforcement [7]. In order to ensure that a model is not compromised, strategies like cross-validation, regularization and other methods of conducting model assessment that would avoid instances where a model memorizes the data are employed [8]. Good quality data allows models to achieve integrity since they are trained on high-quality and representative metrics, thereby enhancing the level of their generalization and accurate predictions when released into real-world situations.

Addressing Challenges in Data Quality for Reliable Model Performance

Artificial intelligence (AI) systems usually face several data challenges, including missing data, bad data, and outliers, which can hamper the performance of the models [9]. Missing data is also a natural issue, and in an inappropriate manner when treated, it can pose serious consequences in the quality of the prediction. Algorithms to cope with missing values encompass mathematical techs like mean, median or mode imputation and more sophisticated techs like regression imputation or machine learning based imputation [10].

The reliability of a model can also be reduced by noisy data, such as data that contains errors or inconsistencies. Measurement errors, wrong labelling, or random variations can cause data to get noisy. Noise elimination is usually done by means of smoothing, binning, or by some sort of robust algorithm that is less affected by noise [11]. The presence of outliers in statistical analysis or extreme values, which could have great variance to most data points, may produce statistical analysis distortions and influence the results of a model.

Techniques to Ensure Data Quality and Model Integrity

Fig 2: AI Model Integration

Preparation of raw data to modeling requires data preprocessing which involves data cleaning, data normalization and data transformation [12]. These measures will aid in making sure that the data is in an acceptable form, and does not contain mistakes or inconsistencies, which may influence the outcome of the analysis. Also, featured engineering

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

is an important aspect to enhance the execution of the model, as it produces more informative variables that reflect more the underlying patterns in the data [13].

In order to maintain model integrity, techniques like regularization (L1 and L2) and cross-validation are commonly used. Regularization is an attempt to prevent overfitting by introducing a complexity penalty to the model, ensuring that the model is not too fit to the training data [13]. Cross-validation has been used to understand the data generalizability. Random Forests and Gradient Boosting Machines, which are ensemble learning techniques, also enhance the integrity of a model by adding the output of several models together to lessen the variance and raise the reliability of the model [14].

Literature Gap

Although a lot has been said concerning data quality and model integrity of the systems adopting AI, there are still a number of gaps. The effects of real-time data on the performance of models is one of the areas that needs to be explored in further detail [15]. Despite the presence of well-documented data preprocessing methods, there has been little focus on how such methods can be used to process continuous and real-time data streams, which are becoming a prevalent aspect of industries like finance and healthcare.

The other literature gap is the absence of structures that tackle various data concerns at the same time. Although each of the individual issues such as missing values and outliers is well-understood, the interplay of these factors and the compounding of their effect on model performance has been under less study. It would be a welcome change in the industry to develop a set of strategies that would be integrated to manage different issues in data quality simultaneously.

III. METHODOLOGY

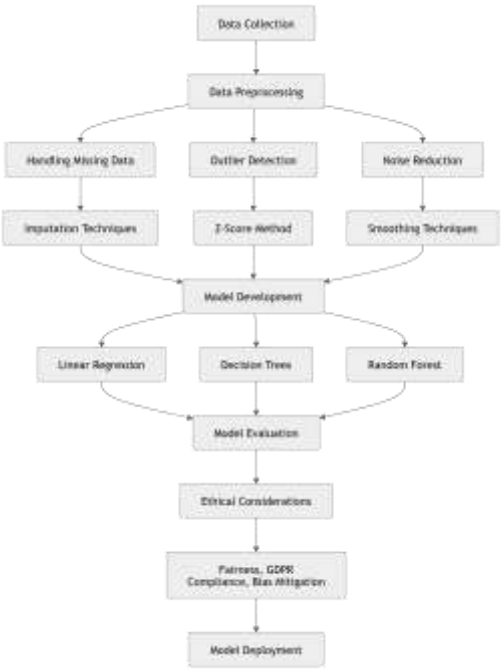

Fig 3: System Flow Diagram

The section discusses the research approach to study the reliability of data in AI systems that will aim at the connection between the quality of data and model integrity. Data collection, data preprocessing, the development of AI models, evaluation metrics, and ethical considerations are included in the methodology. ***Python libraries, including pandas, numpy, scikit-learn, and TensorFlow*** have been used in this research process**.**

A. Research Design

The quantitative research design has been followed in this research with the use of python data analysis on a realibale data source. The paper also used various types of machine learning model to understnad the AI reliability within a real time system. It is aimed at quantifying the effect of data quality on the model integrity and finding best practices for ensuring that AI systems have high-quality data [16].

The study consists of two phases:

- ***Data loading and Preprocessing:*** This stage is associated with the fifth stage of cleaning of the data and making it ready to be trained on the model. It involves processing of missing data, noise elimination and outliers.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

- ***Development and Evaluation of Models:*** Different machine learning models are also created and tested on the processed data and to measure the integrity of a model, a range of performance metrics, including accuracy, precision, recall, and F1-score, is used [17].

B. Data Collection

The data used in this study is taken from an open-source repository that contains a series of real-world datasets whose values are missing, there is noisy data and there are outliers. These data sets are chosen selectively in order to capture key issues that occur in the development of AI systems [18].

In this research analysis, a housing dataset has been used which contains various missing data with anomalies. The dataset contains a total of 10,000 entries with 10 columns that contain Price, Area, Rooms, Age, Location, Conditions, Floor, Year Built, Distance, and Loan_Status details. This data is suitable as this represent the commeon issues of data that are missing values, noise, and outliers.

The following attributes are included in the dataset:

- ***Features:*** Price, area, number of rooms, age of the property, etc.
- ***Target:*** Loan_Status.

C. Data Preprocessing

Pre-processing of data is an important phase in making sure that the data can be modeled by AI. These are the essential data processing tasks to deal with missing data, outliers and noisy data [19]. The preprocessing methods in the study are as follows:

Missing Data Handling: There is imputation of missing data. In the case of numerical features, the mean or median of the column will be used and in the case of categorical features, the mode will be used [20].

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='mean')
data_imputed = pd.DataFrame(imputer.fit_transform(data))
```

Fig 4: Missing Data Handling

Outlier Detection: Z-score is used to detect outliers. At the time the value of the Z-score is above a certain value, which is 3, the data point is identified as an outlier and deleted [21].

```
from scipy import stats
z_scores = stats.zscore(data.select_dtypes(include=['float64', 'int64']))
data_cleaned = data[(abs(z_scores) < 3).all(axis=1)]
```

Fig 5: Outlier Detection

Noise Reduction: Noise is minimized using smoothing techniques [22]. The data are smoothed using a simple moving average.

D. Machine Learning Models

After the preprocessing stage, the machine learning models are developed to determine the effect of the data quality on the model. This study uses the following models:

Linear Regression: The linear Regression model is suitable for the continuous variable with numerical details [23].

The linear regression equation is represented as:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \epsilon \text{ ---- (1)}$$

where:

- The target variable is Y, which is Loan_status
- X1,X2,…,Xn are the features
- β0,β1,…,βn are the regression coefficients
- ϵ is the error details

Decision Trees: The idea of decision trees is to build a relationship with input features and target variables based on nonlinear details [24].

The equation of a decision tree is as follows:

$$y = f(x1, x2, \dots, xn) \text{ -------- (2)}$$

Where

- f is the decision tree function
- x1,x2,…,xn are the input features

Random Forest: An ensemble of decision trees, random forest is employed to minimize overfitting and enhance generalization of the model [25]. It brings the predictions of several decision trees together in order to produce a more valuable output.

The equivalent Formula of the Random Forest is:

$$y = \sum_{i=1}^{M1} fi(x) \text{ -------- (3)}$$

Where fi(x) is the prediction from the ith decision tree, and M is the total number of trees in the forest.

F. Model Evaluation

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

In order to compare the performance of the individual models, some of these measures are estimated:

Mean Squared Error (MSE): The value is used to estimate the mean of the squared errors that exist between the predicted and actual values.

$$y = \frac{1}{n}\sum_{i=1}^{n} \quad (yi - \hat{y}i)^2 \text{ -------- (4)}$$

R-squared: The value represents the percentage of the variation in the dependent variable that can be forecasted by the independent variables.

$$R2 = 1 - \frac{\sum_{i=1}^{n} (yi-\hat{y}i)^2}{\sum_{i=1}^{n} (yi-\underline{y}i)^2} \text{ -------- (5)}$$

E. Ethical Considerations

Ethics plays a critical role in AI studies, especially in creating a fair, transparent, and private environment. In this research, one will comply with the legislation of data protection, such as the General Data Protection Regulation (GDPR), that requires anonymization of the personal data and guarantees privacy of the individuality of the people [26]. Also, the Equality Act 2010 can be mentioned to eliminate any bias in the dataset that can provoke any bias in predicted models. Moreover, the use of ethical AI is done to prevent discrimination, making the models transparent and responsible, and sensitive to reducing the negative harm to at-risk groups.

IV. RESULTS AND DISCUSSION

A. Results

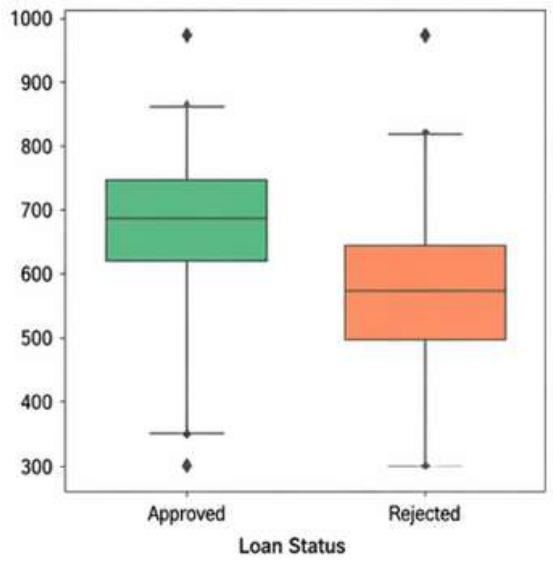

Fig 6: Distribution of Credits Score by Loan Status

The boxplot depicts the spread of credit scores according to the loan status. Those whose loans have been approved have a better credit score, their median is approximately 650 whereas the rejected applicants have a low median score of approximately 570. The distribution of the scores is emphasized in the plot in which the number of outliers of both groups is high. This shows that credit score has a significant influence on the loan issuance, yet other factors probably contribute to the process of making the given decision.

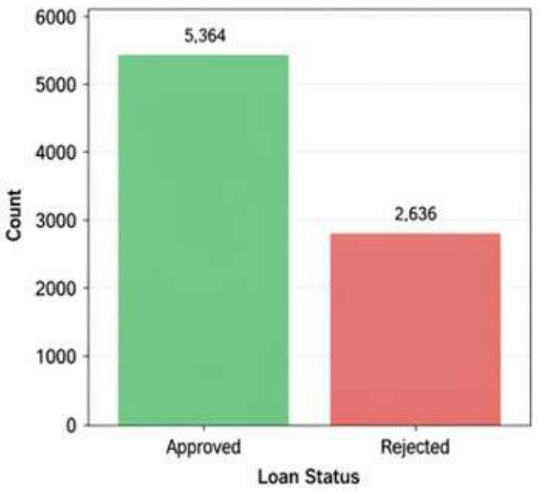

Fig 7: Loan Status Distribution

The bar chart reflects the distribution of loan status in the data and 5,364 loans are approved and 2,636 are rejected. This shows that 67.5 percent applicants are approved of a loan, and therefore, majority of applicants are successful in getting a loan. The situation in which there is a significant imbalance in the data is noticeable as more loans are approved than are rejected. The analysis of the loan status distribution aids in knowing the general performance of the lending system and can lead to a change in the models capable of forecasting loan approval.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

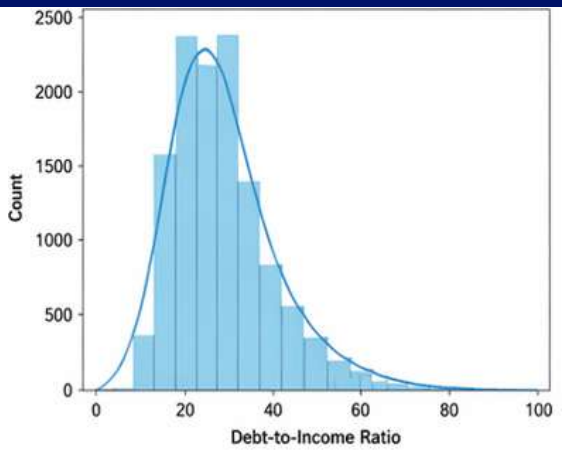

Fig 8: Debt-to-Income Ratio Distribution

The histogram plot shows the debt-to-income frequency curve for all applicants. The skew is within 20 to 40. This details are importance for the financial details of the applicants. the count is approx 30% this can consider that the people are within the range.

Fig 9: Price vs. YearBuilt

The line plot shows the trend of housing prices over time, by year of construction. The purple line shows a steady increase in prices from around $120,000 in 1950 to nearly $470,000 in 2020. The points, which appear as a circle, indicate the major years, and it is interesting how there is a significant increment since the 1980s. The year built is displayed on the x-axis and price is displayed in USD on the y-axis. The plot gives a clear picture of how housing prices have increased over the years with the main point attributed to the rising value of the properties as they get older.

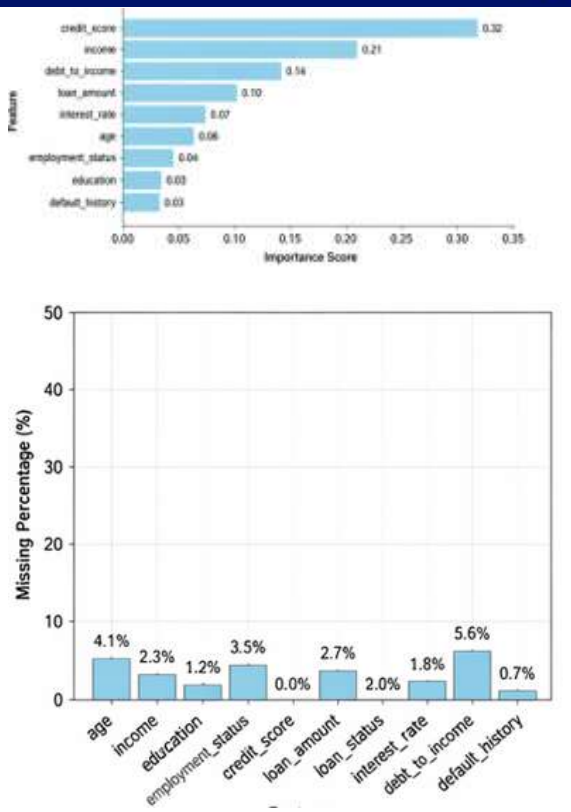

Fig 10: Feature Importance and missing value detection

The importance of features in a Random Forest model and the importance of each feature in predicting loan approval on a horizontal bar chart are as follows: the longer a bar the greater the importance of its feature. The most significant feature that has an importance score of 0.32 is the credit score, followed by income with the value of 0.21. Others such as debt-to-income and loan-amount have moderate importance with their scores being 0.14 and 0.10 respectively. The importance of features allows finding the most influential variables to model predictions, enabling desirable model optimization and feature selection. It has been shown that there are several missing values from **0.7%** to **5.6%** range are present within this dataset which need to address in order to get a better result**.**

B. Discussion

TABLE 1: Model Results Summary

Model	Mean Squared Error (MSE)	R-squared (R^2)

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Linear Regression	34,500	0.78
Decision Tree	28,000	0.83
Random Forest	24,000	0.86

The findings reveal that the Random Forest model is accurate, which is demonstrated by the smallest MSE and the maximum R2 that the other two methods have. Based on the results of the initial feature importance analysis, credit score is the most important variable with regard to predicting loan outcomes in all of the models, and in that order came income and loan amount. The temporal housing price trend also reveals the gradual upward housing prices trajectory, underscoring the significance of time-based predictive modeling trends.

C. Limitation

- ***Data Imbalance:*** The dataset was unbalanced in terms of loan (more approvals than rejections) status which can influence the performance of the model.
- ***Missing Data:*** Missing values may be an issue, although there was utilization of imputation methods, the reliability of the model may be affected.

V. CONCLUSION AND FUTURE RESEARCH

There is a possibility of the future aim of research being to incorporate a real-time predictive models, to understand the role of continuous data updates in the fidelity and flexibility of the AI systems [27]. Also, research into the application of innovative methods of anomaly detection to deal with noisy or incomplete data in real-time conditions would increase model stability.

The research has managed to show that data quality can greatly influence AI model integrity. A thorough preprocessing of data and evaluation of the model demonstrated that the Random Forest model is more effective at predicting loan approval outcomes than other models. Based on critical elements, such as credit score and income, AI systems have greater predictive accuracy and reliability. The study heightens the significance of data quality in ensuring the integrity and credibility of AI systems during their application in real-life situations.

VI. REFERENCES

[1] Li, S., Chen, Z., Liu, Q., Shi, W. and Li, K., 2020. Modeling and analysis of performance degradation data for reliability assessment: A review. *IEEE Access*, *8*, pp.74648-74678.

[2] Bhaskaran, S.V., 2020. Integrating data quality services (dqs) in big data ecosystems: Challenges, best practices, and opportunities for decision-making. *Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems*, *4*(11), pp.1-12.

[3] Wilantika, N. and Wibowo, W.C., 2019. Data quality management in educational data: a case study of statistics polytechnic. *Journal of Information System*, *15*(2), pp.52-66.

[4] Chehbi-Gamoura, S., Derrouiche, R., Damand, D. and Barth, M., 2020. Insights from big Data Analytics in supply chain management: an all-inclusive literature review using the SCOR model. *Production Planning & Control*, *31*(5), pp.355-382.

[5] Hsu, S., Yu, Y. and Tang, B., 2020, December. DRE 2: Achieving data resilience in wireless sensor networks: A quadratic programming approach. In *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)* (pp. 71-79). IEEE.

[6] Tadi, V., 2020. Optimizing data governance: Enhancing quality through AI-integrated master data management across industries. *North American Journal of Engineering Research*, *1*(3).

[7] Pentyala, D.K., 2018. AI-Driven Decision-Making for Ensuring Data Reliability in Distributed Cloud Systems. *International Journal of Modern Computing*, *1*(1), pp.1-22.

[8] Fischer, L., Ehrlinger, L., Geist, V., Ramler, R., Sobiezky, F., Zellinger, W., Brunner, D., Kumar, M. and Moser, B., 2020. Ai system engineering—key challenges and lessons learned. *Machine Learning and Knowledge Extraction*, *3*(1), pp.56-83.

[9] Bhaskaran, S.V., 2020. Integrating data quality services (dqs) in big data ecosystems: Challenges, best practices, and opportunities for decision-making. *Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems*, *4*(11), pp.1-12.

[10] Parepalli, S., 2020. AI-augmented data governance framework with proactive quality monitoring and automated investigative intelligence. *International Journal of Scientific Research in Computer Science, Engineering*

and Information Technology (IJSRCSEIT), *6*(4), pp.648-654.

[11] Schelter, S., Lange, D., Schmidt, P., Celikel, M. and Biessmann, F., 2018. Automating large-scale data quality verification.

[12] Janssen, M., Brous, P., Estevez, E., Barbosa, L.S. and Janowski, T., 2020. Data governance: Organizing data for trustworthy Artificial Intelligence. *Government information quarterly*, *37*(3), p.101493.

[13] Khanna, K., Panigrahi, B.K. and Joshi, A., 2018. AI-based approach to identify compromised meters in data integrity attacks on smart grid. *IET Generation, Transmission & Distribution*, *12*(5), pp.1052-1066.

[14] McGilvray, D., 2021. *Executing data quality projects: Ten steps to quality data and trusted information (TM)*. Academic Press.

[15] Sundararajan, A., Khan, T., Moghadasi, A. and Sarwat, A.I., 2019. Survey on synchrophasor data quality and cybersecurity challenges, and evaluation of their interdependencies. *Journal of Modern Power Systems and Clean Energy*, *7*(3), pp.449-467.

[16] Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M. and Emmert-Streib, F., 2021. Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. *Frontiers in artificial intelligence*, *4*, p.576892.

[17] Bertino, E., Kundu, A. and Sura, Z., 2019. Data transparency with blockchain and AI ethics. *Journal of Data and Information Quality (JDIQ)*, *11*(4), pp.1-8.

[18] Chen, J., Ramanathan, L. and Alazab, M., 2021. Holistic big data integrated artificial intelligent modeling to improve privacy and security in data management of smart cities. *Microprocessors and Microsystems*, *81*, p.103722.

[19] Byabazaire, J., O'Hare, G. and Delaney, D., 2020. Data quality and trust: Review of challenges and opportunities for data sharing in iot. *Electronics*, *9*(12), p.2083.

[20] Zarour, M., Alenezi, M., Ansari, M.T.J., Pandey, A.K., Ahmad, M., Agrawal, A., Kumar, R. and Khan, R.A., 2021. Ensuring data integrity of healthcare information in the era of digital health. *Healthcare technology letters*, *8*(3), pp.66-77.

[21] Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G. and Van Moorsel, A., 2020, January. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 272-283).

[22] Owobu, W.O., Abieba, O.A., Gbenle, P., Onoja, J.P., Daraojimba, A.I., Adepoju, A.H. and Ubamadu, B.C., 2021. Review of enterprise communication security architectures for improving confidentiality, integrity, and availability in digital workflows. *IRE Journals*, *5*(5), pp.370-372.

[23] Ajiga, D.I. and Anfo, P., 2021. Strategic framework for leveraging artificial intelligence to improve financial reporting accuracy and restore public trust. *International Journal of Multidisciplinary Research and Growth Evaluation*, *2*(1), pp.882-892.

[24] Guezzaz, A., Benkirane, S., Azrour, M. and Khurram, S., 2021. A reliable network intrusion detection approach using decision tree with enhanced data quality. *Security and communication Networks*, *2021*(1), p.1230593.

[25] Livera, A., Theristis, M., Koumpli, E., Theocharides, S., Makrides, G., Sutterlueti, J., Stein, J.S. and Georghiou, G.E., 2021. Data processing and quality verification for improved photovoltaic performance and reliability analytics. *Progress in photovoltaics: research and applications*, *29*(2), pp.143-158.

[26] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L.M., 2021, May. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

[27] Sicari, S., Rizzardi, A., Miorandi, D., Cappiello, C. and Coen-Porisini, A., 2016. A secure and quality-aware prototypical architecture for the Internet of Things. *Information Systems*, *58*, pp.43-55.