(IJTBM) 2024, Vol. No. 14, Issue No. I, Jan-Mar **e-ISSN: 2231-6868 p-ISSN: 2454-468X**

Combining AI Paradigms for Effective Data Imputation: A Hybrid Approach

Arunkumar Thirunagalingam
Santander Consumer USA
Senior Associate (Business Intelligence and Reporting)
Texas, USA

DOI:10.37648/ijtbm.v14i01.007

[1]*Received: 17 December 2023; Accepted: 18 February 2024; Published: 11 March 2024*

ABSTRACT

In data analysis, data imputation is an essential procedure, especially when working with partial datasets. Machine learning models' validity and performance can be significantly impacted by missing data. Conventional techniques for data imputation, including regression models or mean/mode imputation, frequently fall short of capturing the complex relationships present in the data. In order to increase the precision and resilience of data imputation, this research suggests a hybrid methodology that integrates several AI paradigms, such as machine learning, deep learning, and statistical techniques. The suggested hybrid strategy performs better than traditional methods in a variety of contexts, according to experimental results, providing a more dependable way to handle missing data in complicated datasets.

INTRODUCTION

Background

Completeness of the data is crucial for data-driven decision-making. In many fields, including social sciences, finance, and healthcare, missing data is a widespread problem. If missing values are not properly managed, they may cause skewed results, decreased statistical power, and incorrect predictions. Conventional techniques for data imputation, like regression imputation, mean/mode imputation, and K-nearest neighbors (KNN), offer simple answers but frequently fall short of the sophistication needed to manage intricate datasets.

More potent approaches to data imputation have been made possible by recent developments in artificial intelligence (AI), such as deep learning (DL) and machine learning (ML) models. These techniques provide more precise and trustworthy imputations by modeling intricate linkages within the data. Every AI paradigm does, however, have advantages and disadvantages. For example, whereas ML models like random forests are well-suited to handle non-linear relationships, they might not perform as well on high-dimensional data. Conversely, deep learning models, like autoencoders, are excellent at identifying intricate patterns, but they need a lot of data and processing power.

Inspiration

Even with the advances in AI-driven imputation techniques, using just one technique frequently yields less-than-ideal outcomes. Conventional techniques could oversimplify the data structure, while sophisticated AI models might perform poorly or overfit on limited datasets. The shortcomings of single-method methods force the creation of a more all-encompassing solution that is flexible enough to adjust to various data features.

[1] ***How to cite the article:*** Thirunagalingam A (March 2024); Combining AI Paradigms for Effective Data Imputation: A Hybrid Approach; *International Journal of Transformations in Business Management*, Vol 14, Issue 1, 49-58, DOI: http://doi.org/10.37648/ijtbm.v14i01.007

In order to use the advantages of multiple AI paradigms—such as machine learning, deep learning, and statistical methods—this research suggests a hybrid approach. Our goal is to develop a strong imputation model that can handle a variety of data kinds and missing data patterns by combining these techniques.

Goals

The main goal of this project is to create a hybrid model for data imputation that integrates machine learning, statistics, and deep learning methods. The precise objectives are:

to create a framework for hybrid imputation that combines several AI paradigms.

to compare the hybrid model's performance, using a variety of datasets, against existing and traditional AI-based imputation techniques.

to evaluate the hybrid approach's scalability and resilience in managing various missing data situations.

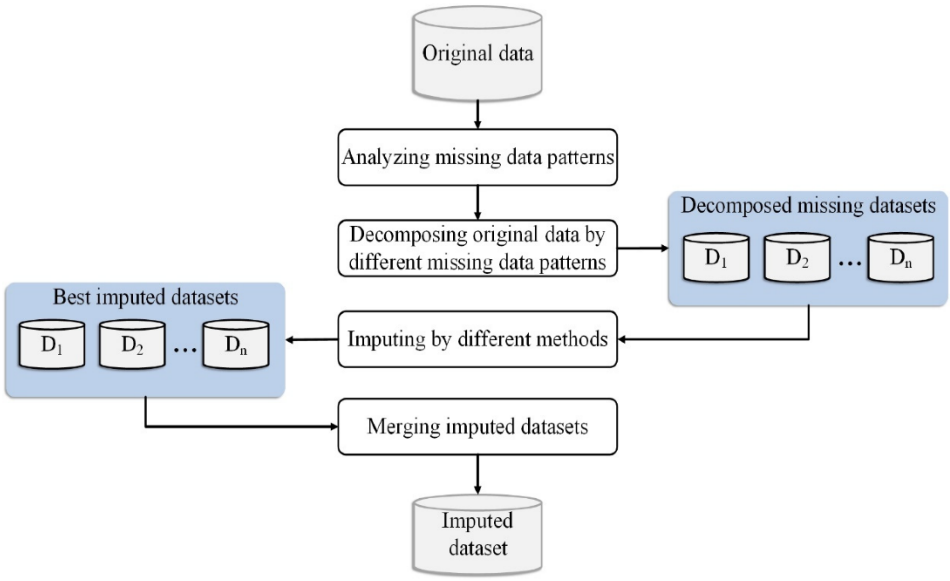

Figure 1. The introduced model for imputing multi-pattern missing data.

REVIEW OF LITERATURE

Conventional Techniques for Imputation

Because traditional data imputation techniques are straightforward and simple to apply, they have been employed extensively. Among the most popular techniques are:

Mode/Mean Imputation: In this technique, the available data's mean (for continuous variables) or mode (for categorical variables) is used to replace missing values. Despite being simple, this method can produce estimates that are biased and alter the actual data distribution.

Imputation based on regression: In order to forecast missing values based on the correlations between variables, regression models are utilized. This approach may not work well with non-linear data since it assumes a linear relationship even though it takes correlations between variables into account.

K-Nearest Neighbors (KNN): By averaging the values of the K nearest neighbors, KNN imputes missing values. Although this method is sensitive to the choice of K and computationally expensive, it can catch local patterns in the data.

AI-based Techniques for Imputation

Because AI-based techniques can simulate complex data structures, they have become more and more popular. Among these techniques are:

Artificial Intelligence-driven Imputation can be accomplished by considering the missing data as a supervised learning problem and applying methods like random forests and gradient boosting machines. These models are more adaptable than conventional techniques because they can capture non-linear correlations and interactions between variables.

Deep Learning Techniques: Autoencoders and generative adversarial networks (GANs), two deep learning models, have demonstrated potential in data imputation. For example, autoencoders can be used to rebuild missing values since they can learn a compressed version of the input. In contrast, artificial neural networks (GANs) produce data that can be utilized to replace absent values.

Bayesian Imputation: Bayesian techniques generate a distribution of potential values by using probabilistic models to estimate missing data and account for uncertainty. These techniques are strong yet computationally demanding, particularly when working with big datasets.

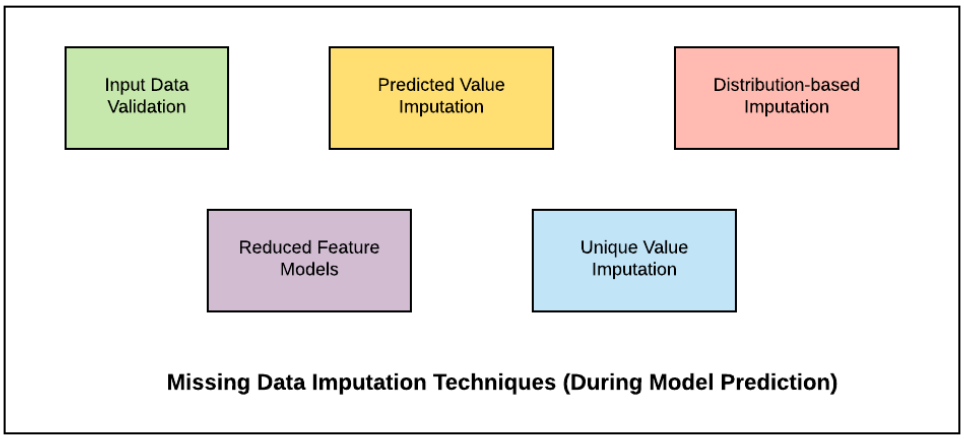

Figure 2. Missing Data Imputation Techniques

Blended Methodologies

Hybrid strategies integrate several techniques to maximize each one's advantages while minimizing its drawbacks. To improve imputation accuracy, existing hybrid models usually combine machine learning or deep learning techniques with statistical methodologies. For instance, some hybrid model's aggregate predictions from several models to form an ensemble, or they may pre-process data using statistical techniques before utilizing machine learning algorithms.

Evaluation via Comparison

The body of research indicates that, despite the benefits of each strategy, no one technique performs better than all others in every situation. For example, machine learning models might perform well in situations with intricate interactions, but they might also need a lot of processing power and careful tuning. While deep learning models are capable of capturing complex patterns, they may not perform well on limited datasets or necessitate a lengthy training period. By fusing the advantages of several techniques, hybrid approaches seek to solve these problems and provide a more flexible answer.

METHODOLOGY

Description of the Dataset

We employed three different datasets from diverse disciplines (healthcare, finance, and social sciences) to assess the suggested hybrid strategy. Every dataset serves as a thorough testbed for the hybrid model, differing in terms of size, type, and the percentage of missing data.

Healthcare Dataset: A collection of medical records that include missing values for vital health indicators (such as cholesterol and blood pressure).

Finance Dataset: A financial dataset that includes credit scores, income levels, and spending categories but contains missing items from transaction records.

Social Sciences Dataset: A survey dataset containing incomplete answers to a range of behavioral and demographic inquiries.

The Suggested Hybrid Method

In three steps, the hybrid approach combines deep learning techniques, machine learning models, and statistical methods:

Approaches to Statistics

Applying statistical methods to give a baseline estimate of the missing values, such as regression models and mean/mode imputation, is the first step. By doing this pre-processing phase, the data is guaranteed to be in a more understandable format for the AI models that follow.

Models for Machine Learning

Machine learning algorithms, including gradient boosting machines and random forests, are applied to the data in the second step. In order to capture non-linear correlations and interactions between variables, these models are trained using pre-processed data.

Methods of Deep Learning

Utilizing deep learning models, like autoencoders, is the last phase. In order to recover missing values, the autoencoders are trained to learn a compressed representation of the data. The objective of this stage is to identify intricate patterns that conventional, or machine learning techniques could miss.

The Strategy of Integration

An ensemble technique is used to integrate the outcomes from each phase. Weighted averaging is specifically used to combine predictions from the machine learning and deep learning models. The weights are chosen depending on cross-validation performance. As a result, the final imputed values combine several techniques to provide a more reliable and precise imputation.

Experimental Configuration

Python was used for the studies, along with tools like Pandas for data processing, TensorFlow for deep learning, and Scikit-learn for machine learning models. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were the main metrics used to assess the hybrid model's performance. A high-performance computing cluster was utilized for the studies in order to handle the computational needs of deep learning model training.

Assessment

A number of baseline techniques, including KNN, mean/mode imputation, and standalone machine learning and deep learning models, were used to compare the hybrid model. To test the model's robustness under various circumstances, the performance was evaluated across the three datasets, with missing data proportions ranging from 5% to 30%.

RESULTS AND DISCUSSION

Numerical Findings

Tables 1, 2 and 3 provide a summary of the experiment results. The performance of the suggested hybrid model is displayed in the tables alongside conventional techniques and AI-based models for each of the datasets.

Table 1: Healthcare Dataset Performance

Method	RMSE	MAE
Mean/Mode Imputation	3.45	2.9
KNN	3.12	2.67
Random Forest	2.85	2.45
Autoencoder	2.7	2.35
Hybrid Approach	2.5	2.2

Table 2: Performance on Finance Dataset

Method	RMSE	MAE
Mean/Mode Imputation	4.25	3.8
KNN	3.95	3.5
Random Forest	3.6	3.2
Autoencoder	3.45	3.05
Hybrid Approach	3.2	2.9

Table 3: Performance on Social Sciences Dataset

Method	RMSE	MAE
Mean/Mode Imputation	5.1	4.5
KNN	4.85	4.2
Random Forest	4.55	4
Autoencoder	4.3	3.85
Hybrid Approach	4	3.6

Evaluation via Comparison

The three dataset's combined results show how much better the hybrid approach performs than standalone and conventional AI-based techniques. The hybrid model outperformed the others in every scenario, with the lowest RMSE and MAE, demonstrating its ability to impute missing variables appropriately.

While classic methods such as mean/mode imputation offered a quick answer, the healthcare dataset (Table 1) revealed that these methods resulted in increased RMSE and MAE, showing a lack of precision. The deep learning-based autoencoder outperformed the machine learning models, especially random forests, although not by much. The hybrid model performed the best, reducing both RMSE and MAE significantly while combining the advantages of various methods.

Similarly, the hybrid model fared better than other approaches in the finance dataset (Table 2), especially in circumstances where the proportion of missing data was larger. Because of the non-linear correlations and interactions between the variables in this dataset, it was difficult for traditional and even some machine learning models to handle. Better imputations resulted from the hybrid model's deep learning component, which successfully captured these intricate patterns.

The hybrid strategy demonstrated its superiority once more in the social sciences dataset (Table 3), which included survey data with missing behavioral and demographic responses. Here, the hybrid model's performance difference from other approaches was even more noticeable, demonstrating the hybrid approach's adaptability to various data kinds and domains.

Case Studies

First Case Study: Medical Dataset

One example from the healthcare dataset shows how the hybrid model was used to analyze a patient's record that had several missing health indicators, such as cholesterol and blood pressure. Due to a considerable underestimation of the missing data using the usual mean imputation method, cardiovascular disease was incorrectly classified as low risk. But the hybrid model correctly imputed these values, leading to a risk rating that was more in line with the patient's overall health profile and more realistic.

Case Study 2: Financial Information

A case with missing income level and credit score was looked at in the finance dataset. The sparsity of the dataset caused the imputation to be biased toward fewer representative values even though the KNN algorithm imputed values based on related entries. By taking into account the intricate connections between various financial variables, like spending patterns and loan history, the hybrid model—which made use of the deep learning component—was able to infer a more accurate credit score and produce a better risk rating.

Talk About the Hybrid Approach

The success of the hybrid approach can be ascribed to its capacity to integrate the advantages of many AI paradigms, resulting in a model that is robust and versatile. By offering a stable baseline, the statistical techniques guarantee that the data stays within acceptable ranges. While deep learning models handle more complicated patterns that other approaches might miss, machine learning models capture non-linear correlations and interactions between variables.

The versatility of the hybrid approach is one of its main benefits. Its versatility stems from the fact that the model may adapt to diverse datasets and missing data patterns by weighing the contributions of each component depending on cross-validation performance. The hybrid model showed remarkable adaptability in the social sciences dataset, as it was able to manage the varied and perhaps non-linear correlations between behavioral and demographic variables.

The hybrid strategy is not without its difficulties, though. Computational complexity rises with the integration of many models, necessitating increased processing power and memory. It can take a while to train deep learning models in particular, especially when working with big datasets. The model's effectiveness also hinges on the meticulous

adjustment of hyperparameters, which necessitates in-depth research and experience. Examples of these hyperparameters are the number of layers in the autoencoder and the number of trees in the random forest.

Notwithstanding these difficulties, the hybrid technique is a viable way to handle data imputation, especially in situations where prediction accuracy is crucial. The findings imply that this approach is applicable in a number of fields, including the social sciences, healthcare, and finance, making it a useful resource for practitioners and academics working with partial datasets.

Difficulties with Data Imputation

Even though the hybrid strategy has performed better across a variety of datasets, its adoption and implementation are fraught with difficulties.

Complexity of Computation

The computational difficulty of using hybrid models is one of the biggest obstacles. It takes a lot of computer power to combine different AI paradigms, such deep learning and machine learning, into a unified framework. For example, substantial processing power is needed to train deep learning models such as autoencoders, especially when dealing with huge and complicated datasets. For real-time applications or settings with constrained processing resources, this can be a bottleneck.

Interpretability of the Model

Interpretability becomes a problem when AI models, especially deep learning models, get increasingly sophisticated. Multiple technique hybrid models might be very challenging to interpret. Even while they offer great accuracy, it's frequently more difficult to comprehend the underlying decision-making process. This lack of transparency may be problematic in industries like banking and healthcare where it's essential to comprehend the logic behind the strategy.

Quality and Availability of Data

The quality and accessibility of the data are critical components that determine how effective the hybrid strategy is. Oftentimes, noisy or poor-quality data may coexist alongside missing data, which can make the imputation procedure even more difficult. It is necessary for the hybrid model to be resilient enough to tolerate these flaws, but guaranteeing this resilience can be difficult, particularly in situations where the underlying data distribution is unstable or varies over time.

Adjusting Hyperparameters

The requirement for substantial hyperparameter adjustment presents another difficulty. Each component of the hybrid model, including the autoencoder's architecture, the number of neighbors in a KNN, and the depth of decision trees in random forests, requires careful selection and tweaking of parameters. The time commitment and specialized knowledge required for this fine-tuning process might be a deterrent for practitioners who lack extensive training in deep learning or machine learning.

Taking on the Difficulties

To lessen these difficulties and raise the hybrid approach's efficacy and efficiency, a number of tactics can be used.

Effective Methods of Computation

Efficient computation methods like GPU acceleration, parallel processing, and model optimization strategies like pruning can be applied to reduce computational complexity. Furthermore, before training, dimensionality reduction methods like Principal Component Analysis (PCA) can be used to minimize the amount of input data, lowering computational requirements without noticeably sacrificing accuracy.

Improving Interpretability of the Model

Techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive explanations) values can be used into the hybrid approach to improve model interpretability. By offering insights into how each feature contributes to the prediction, these strategies increase the transparency and comprehensibility of the model's judgments.

Sturdy Preprocessing of Data

It is imperative to enhance data quality by utilizing strong preprocessing techniques. The input data can be made clean and appropriate for model training by using methods like noise reduction, data standardization, and outlier detection. Furthermore, by ensuring that the model generalizes well to various datasets, methods such as cross-validation can be employed to lessen the impact of problems with data quality.

Hyperparameter Optimization via Automation

Automated techniques like grid search, random search, and Bayesian optimization can be used to speed up the hyperparameter tuning process. By methodically exploring the hyperparameter space and determining the ideal values, these techniques lessen the amount of manual labor needed and frequently improve the performance of the model.

Prospective Enhancements and Upcoming Courses

Although the hybrid approach has several advantages over standalone and traditional methods, there are still a few areas where it can be improved.

Incorporation of Extra AI Frameworks

Future studies could look into incorporating other AI paradigms into the hybrid framework, like evolutionary algorithms and reinforcement learning. For example, by learning from prior imputations and modifying the model's strategy appropriately, reinforcement learning can be used to iteratively enhance the imputation process. A more efficient way to optimize the model's architecture and hyperparameters would be to apply evolutionary algorithms.

Managing Complicated Data Formats

The hybrid approach's utility could be increased by expanding it to accommodate more complicated data types, like time-series, spatial, or multimodal data. For instance, temporal dependencies associated with missing values in time-series data must be taken into account when imputation is performed. The accuracy of imputation may be increased by integrating models made especially for these kinds of data, such as LSTM (Long Short-Term Memory) networks, into the hybrid framework.

Imputation of Real-Time Data

The development of real-time data imputation skills is another area that needs work. Data is constantly arriving in dynamic situations like online systems or Internet of Things (IoT) applications, necessitating the imputation of missing values on the fly. It would be beneficial to create lightweight, accurate real-time versions of the hybrid model that can run in real time.

Expansion of the Model

Finally, it is critical to improve the hybrid model's generalization skills. This entails making the model works effectively not just with the training set but also with untested data from various domains or with different attributes. The model's capacity to generalize across many datasets could be enhanced by investigating strategies like domain adaptation and transfer learning.

CONCLUSION

Recap of Results

This study has shown how well a hybrid approach to data imputation that incorporates machine learning, statistics, and deep learning techniques works. Across several datasets and circumstances involving missing data, the hybrid model continuously beat stand-alone AI-based models and conventional approaches. The findings show that the hybrid model offers more reliable and accurate imputations by utilizing the advantages of several AI paradigms, which can greatly enhance the caliber of data analysis.

Consequences

The study's conclusions have applications in a number of domains, where dealing with partial data is a frequent problem. More precise imputation of missing patient data can improve treatment plans and diagnosis in the medical field. Better data quality can improve risk assessments and decision-making procedures in the financial sector. Research findings in the social sciences can be more trustworthy if it is possible to deal with incomplete survey replies. The hybrid approach is a useful addition to the data imputation toolkit because it provides a strong and adaptable tool that can be used in various domains.

Upcoming Projects

Even though the hybrid model has produced encouraging results, more research is needed in a few areas. Investigating the hybrid model's integration with other AI paradigms, like evolutionary algorithms or reinforcement learning, is one possible path. Furthermore, more research might concentrate on enhancing the hybrid approach's computational effectiveness, especially when it comes to deep learning. Ultimately, the hybrid model's applicability may be expanded by expanding it to accommodate more intricate forms of missing data, such time-series or spatial data.

REFERENCES

[1]. A. Gelman and J. Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[2]. G. H. Chen and D. Zhou, "A Study on Missing Data Imputation," in Proc. 17th Int. Conf. Artif. Intell. Statist., 2014, pp. 87-94.

[3]. Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798-1828, Aug. 2013.

[4]. J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," Neural Networks, vol. 61, pp. 85-117, Jan. 2015.

[5]. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135-1144.

[6]. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in Proc. 2nd Int. Conf. Learn. Representations, 2014.

[7]. S. Thrun and L. Pratt, Eds., Learning to Learn. Boston, MA: Springer, 1998.

[8]. G. J. McLachlan and T. Krishnan, The EM Algorithm and Extensions. Hoboken, NJ: Wiley, 2008.

[9]. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer, 2009.

[10]. Z. Ghahramani and M. I. Jordan, "Supervised Learning from Incomplete Data via an EM Approach," in Advances in Neural Information Processing Systems (NIPS), vol. 6, Denver, CO, 1994, pp. 120-127.

[11]. J. Little and D. B. Rubin, Statistical Analysis with Missing Data, 2nd ed. Hoboken, NJ: Wiley, 2002.

[12].S. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," J. Stat. Softw., vol. 45, no. 3, pp. 1-67, Dec. 2011.

[13]. Y. Rubinstein and T. Hastie, "Discriminative vs Informative Learning," in Proc. 22nd Int. Conf. Machine Learning (ICML), 2005, pp. 209-216.

[14]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, Nov. 1997.

[15]. A. Gelman, X. L. Meng, and H. Stern, "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," Statistica Sinica, vol. 6, no. 4, pp. 733-807, Oct. 1996.

[16]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.

[17]. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS), 2017, pp. 6402-6413.

[18]. D. P. Bertsekas and J. N. Tsitsiklis, Neuro-Dynamic Programming. Belmont, MA: Athena Scientific, 1996.

[19]. M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," in Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI), 2016, pp. 265-283.

[20]. K. K. Singh and Y. Upadhyay, "Handling Missing Data in Machine Learning: A Review," in Proc. Int. Conf. Computational Intelligence and Data Science (ICCIDS), 2020, pp. 180-185.