Early Autism Disorder Prediction Using Machine Learning

Ahmad Ridlan, Muhaimin Hasanudin, Ofelia Cizela da Costa Tavares, Daniel Eliazar Latumaerissa

***Abstract*— Autism is a behavioral disorder caused by neurodevelopmental disorders in the brain. This condition makes it difficult to communicate, socialize, and learn. One of the machine learning algorithms used for early autism prediction is the C4.5 algorithm, which considers the importance of attributes in effectively dividing data. Linear regression is used to predict early autism and identify associations between certain attributes and the likelihood of autism. The aim of this research is to develop a predictive model for early detection of autism using the C4.5 algorithm and linear regression. Additionally, the research aims to address the need for accurate and timely predictions of autism, as the long and expensive diagnostic process causes delays in intervention. This research emphasizes the importance of early detection in the child's growth and development process for rapid intervention and treatment to improve the quality of life for individuals with autism. The dataset from the Machine Learning Repository consists of 1122 data instances with 20 attributes. The results show that the C4.5 algorithm achieved the highest accuracy of 94%, while the linear regression algorithm achieved 44%. These findings suggest that the C4.5 algorithm is more effective in predicting autism than linear regression.**

***Keywords*— Prediction, Autism, Linear regression, Machine Learning.**

I. INTRODUCTION

This research explores the application of machine learning techniques for predicting autism spectrum disorder (ASD) at an early stage [1,2]. The study aims to utilize the linear regression algorithm and the C4.5 algorithm to develop predictive models for identifying ASD in individuals [3,4,5]. By leveraging these algorithms, the research seeks to enhance the accuracy and efficiency of early diagnosis of ASD, which is crucial for timely intervention and support for affected individuals. In the realm of autism prediction, machine learning algorithms play a crucial role. Various studies have investigated the use of different machine learning techniques, such as supervised, unsupervised, and reinforcement learning, to predict autism behaviors and diagnose ASD. These algorithms provide a data-driven approach that can analyze patterns and features to make precise predictions regarding autism spectrum disorder [6-8].

Additionally, the research aligns with prior studies that have utilized data mining algorithms like the C4.5 decision tree[9], linear discriminant analysis (LDA), and extreme learning machine (ELM) to predict early chronic kidney disease (CKD) and subgroups of autism spectrum disorder based on diagnostic criteria. These algorithms have shown effectiveness in predictive modeling for medical conditions, demonstrating their potential in aiding early diagnosis and intervention [10].

Furthermore, the study is consistent with research that applies artificial intelligence and machine learning for the diagnostic classification of autism spectrum disorder. By employing machine learning models, researchers aim to classify and predict ASD subgroups based on specific criteria, such as the Diagnostic and Statistical Manual of Mental Disorders (DSM), underscoring the utility of these techniques in healthcare and disorder prediction [11, 12].

The aim research is to use the C4.5 algorithm and linear regression to achieve the highest accuracy in predicting autism, then compare the two algorithms to improve the early autism disorder data. To predict has the best accuracy value. To continue the comparison with several other algorithms carried out in previous research, which are useful for comparison. This dataset was obtained from a machine learning repository entitled Autism Screening Adult Dataset and contains a total of 122 data and 20 attributes. The highest accuracy result is 94% for the C4.5 algorithm and 44% for the linear regression algorithm

II. RESEARCH METHOD

The data for this study is obtained through a website called the Machine Learning Repository, accessible via the link https://archive.ics.uci.edu/ specifically from the "Autism Screening Adult Data Set." The dataset consists of 1122 instances and 20 attributes, stored in either .xlsx or .csv format. The obtained data include information about autism disorders, symptoms, and decisions related to autism disorders in children. Machine learning is a subfield of artificial intelligence that involves computer programming to achieve a level of intelligence similar to humans and enhance its understanding through automatic learning from experiences [13,14]. The importance of machine learning lies in developing systems that can learn autonomously to make decisions without being repeatedly programmed by humans. This enables machines to not only make decisions but also adapt to changes that occur. In practice, machine learning works by analyzing large datasets (big data) as input, with the aim of discovering patterns within them. This data is used as input to train the machine, enabling it to produce accurate analyses. In the context of machine learning, there are the concepts of training data and testing data. Training data is used to train algorithms in machine

Manuscript received October 05, 2024.

learning, while testing data is used to evaluate the performance of the trained algorithms when faced with new data that was not provided during the training process [15].

The C4.5 algorithm is derived from the ID3 algorithm developed by Ross Quinlan. The C4.5 algorithm is used to categorize information that has numeric or categorical attributes. The result of the classification process in the form of rules can be utilized to predict the values of categorical attributes for new entries [16,17]. In this research, the C4.5 algorithm is performed by inputting the data, followed by preprocessing. The preprocessing phase includes actions such as data cleaning, data transformation, and data splitting. The preprocessing process is aimed at facilitating the modeling process for prediction. After the preprocessing stage, the C4.5 modeling is carried out. In this modeling step, the decision tree from C4.5 is used to assess the accuracy of the algorithm and visualize the results. Linear Regression is a statistical method used to model the linear relationship between independent variables (predictor variables) and a dependent variable (target variable). The main objective of linear regression is to find the best-fitting straight line (regression line) that represents the linear relationship between the variables. Linear Regression is also considered a simple yet effective method for modeling the relationship between a dependent variable and one or more independent variables. This method assumes a linear relationship between the variables, meaning that the relationship can be represented by a straight line [18].

In this research leverages data encompassing information about autism disorders and symptoms to develop predictive models using machine learning techniques, particularly the C4.5 algorithm and Linear Regression. The integration of this data is essential for training and testing the machine learning algorithms, ultimately contributing to the improvement of automated detection capabilities for early autism prediction. The C4.5 algorithm and Linear Regression, utilizes training and testing data, and integrates obtained data to develop predictive models for early autism detection. The Linear Regression algorithm involves inputting the data, followed by preprocessing. The preprocessing phase includes actions such as data cleaning, data transformation, and data splitting. The purpose of the preprocessing process is to facilitate the modeling process for prediction. After preprocessing, the Linear Regression modeling is conducted. In this modeling step, RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) from Linear Regression are used to assess the accuracy of the algorithm and visualize the results. The inclusion of RMSE and MAE values is necessary because for Linear Regression, the general form of accuracy results is derived from RMSE and MSE (Mean Squared Error). Overall, the supporting theories of Linear Regression and C4.5 provide a conceptual foundation for using these methods in predicting accuracy. By understanding the basic principles of both methods, researchers can apply them effectively and consider various factors that influence the results. There are several stages involved in this study. These stages are depicted in Figure 1.

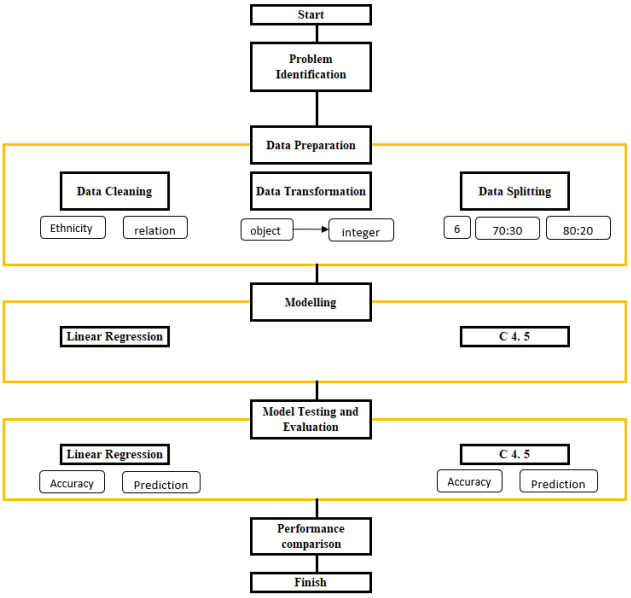


Figure 1 Proposed Model

III. DATASET

Identification data can be defined as a category of data that contains information used to identify or distinguish a specific individual or entity. The dataset of "Autism Screening Adult" contains 20 features listed in Table 1 below.

Table 1. Variables Table of the dataset

Variable Name	Role	Type
A1_Score to A10_Score (10)	Feature	Integer
Age		Categorical
Gender		Categorical
Ethnicity		Categorical
Jaundice		Categorical
Family_pdd		Categorical
Country_of_res		Categorical
Used_app_before		Categorical
Result		Integer
Age_desc		Categorical
Relation		Categorical
Class	Target	Categorical

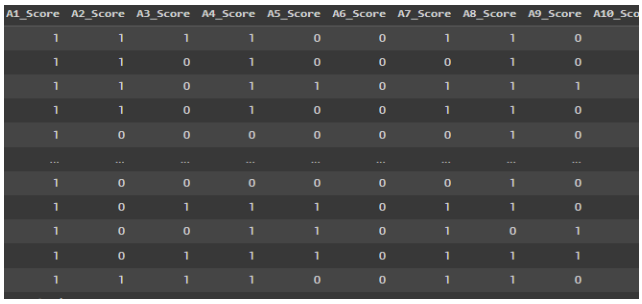
IV. DATA PREPROCESSING

Data pre-processing is a stage where data cleaning, data transformation and data splitting were performed to eliminate irrelevant columns/data and convert data from character to numeric to facilitate the computational process. Data Transformation changes the data type of each attribute using the LabelEncoder method from the sklearn.preprocessing library as shown in the script below.

```
from sklearn.preprocessing import LabelEncoder
encoder = LabelEncoder ()
for col in ['gender', 'ethnicity', 'jundice', 'austim', 'contry_of_res', 'used_app_before', 'age_desc', 'relation', 'Class/ASD']:
```

```
autism_data[col] =
encoder.fit_transform(autism_data[col])
```

Transform loop will iterate through a list of column names: gender, ethnicity, jundice, austim, country_of_res, used_app_before, age_desc, relation, Class/ASD. For each column, the code in the python script will apply the fit_transform() method of the encoder object to convert categorical values into numerical labels. The fit_transform() method will transform the encoder on the column data and convert it to a numeric label. The numeric labels are then reassigned to the corresponding columns in the data. The result of the python script is that the categorical variables in the specified columns in the data are replaced with encoded numeric labels. This stage is useful to facilitate the next stage, to provide a more detailed description of the dataset that has been transfotm, as shown in Figure 3.



A1_Score	A2_Score	A3_Score	A4_Score	A5_Score	A6_Score	A7_Score	A8_Score	A9_Score	A10_Sco
1	1	1	1	0	0	1	1	0	
1	1	0	1	0	0	0	1	0	
1	1	0	1	1	0	1	1	1	
1	1	0	1	0	0	1	1	0	
1	0	0	0	0	0	0	1	0	
...	...	...	...	...	...	...	...	...	
1	0	0	0	0	0	0	1	0	
1	0	1	1	1	0	1	1	0	
1	0	0	1	1	0	1	0	1	
1	0	1	1	1	0	1	1	1	
1	1	1	1	0	0	1	1	0	

Figure 2 Raw Data

The next stage after Data Transformation is carried out, the data is still not ready to be predicted and must pass the Split Data stage. In this research, Split Data is carried out into several categories, namely 70:30, 80:20, and 60:40. Split Data is done by entering the train_test_split method from the sklearn.model_selection library. The source code for the Split Data process can be seen in the script below:

```
from sklearn.model_selection import train_test_split
X = autism_data.drop("Class/ASD", axis=1)  # Features
(input)
y = autism_data["Class/ASD"]  # Target variable
(output)
# Split the data into training and testing datasets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3)  # Split Category Data(0.2; 0.3; 0.4)
```

From the three categories of Split Data above, namely 70:30 (70 data_training and 30 data_testing), the second category is 80:20 (80 data_training and 20 data_testing), and the last category is 60:40 (60 data_training and 40 data_testing), after running the split data script, the results are obtained as in Table 2.

Table 2. Result of Data Splitting

Category	Train Set (rows)	Power (rows)
80:20	897	225
70:30	785	337
60:40	673	449

V. MODELING

In this research, there are two models used for modeling: Linear Regression and C4.5. These models are used to compare the prediction results from each model.

A. *Linear Regression (LR)*

The first modeling technique is Linear Regression, which is used to predict using the scikit-learn library. The source code for modeling Linear Regression can be seen below.

```
lr = LinearRegression()
lr.fit(X_train, y_train)

# Use the trained linear regression model to predict the
labels of the test set
y_pred_lr = lr.predict(X_test)
```

This line creates an instance of the LinearRegression class, representing the linear regression model. This line trains the linear regression model (lr) using the training data. It uses the fit() method, which takes two arguments: X_train, representing the features or independent variables, and y_train, representing the target or dependent variable. This line uses the trained linear regression model (lr) to predict labels on the test dataset. It applies the predict() method on the test dataset (X_test), and the predicted labels are stored in the y_pred_lr variable.

B. *C4. 5*

The second modeling technique is C4.5 which is used to predict using the scikit-learn library. The source code for modeling C4.5 can be seen below.

```
# Train a C4.5 decision tree classifier on the training set
clf = DecisionTreeClassifier(criterion="entropy")
clf.fit(X_train, y_train)

# Use the trained decision tree classifier to predict the
labels of the test set
y_pred_c45 = clf.predict(X_test)
```

An instance of the DecisionTreeClassifier class from scikit-learn is created. The "criterion" parameter is set to "entropy", indicating that the decision tree will use entropy as the criterion to measure the quality of splits. Then, the decision tree classifier is trained using the fit() method. This method takes two arguments: X_train and y_train. X_train represents the feature set of the training data, which is the input variables used for making predictions. y_train represents the corresponding labels or target values for the training set. Once the decision tree classifier is trained, the code proceeds to predict labels for the test set using the predict() method. This method takes the feature set of the test data (X_test) as input and returns the predicted labels (y_pred_c45).

VI. RESULTS AND ANALYSIS

After modeling is done, the next step is data visualization which is useful for visual representation of the

relationship between pairs of variables in the dataset using the seaborn library with the pairplot() function. This can help identify patterns, correlations, and distributions among selected columns, thus enabling a better understanding of the data. Research results in the form of data visualization and model comparison tables with multiple data splits.

A. Visualization

In this stage, researchers conduct data visualization to gain a better understanding of the characteristics and patterns present in the dataset. Data visualization helps researchers identify relationships between the existing attributes and observe potential trends. In this research, researchers utilize various visualization methods such as graphs, plots, and diagrams to represent the data intuitively and in an easily understandable manner. Some examples of visualizations that can be used include scatter plots, bar plots for linear regression, and decision trees for C4.5.

Through data visualization, researchers can observe the distribution of attributes relevant to house prices, such as land area, number of bedrooms, or property age. Researchers can also identify patterns or correlations between these attributes in the " Autistic Spectrum Disorder Screening Data for Children " data. Data visualization also aids researchers in identifying outliers that may exist in the dataset. Outliers can have a significant impact on prediction outcomes and need to be considered in the analysis process. By visualizing the data, researchers can gain deeper insights into the characteristics of the dataset and validate existing assumptions or hypotheses. This helps researchers understand the data better and prepare for the next steps in the analysis and modeling process. By combining visual analysis with statistical methods and machine learning, researchers can develop a more comprehensive understanding of the dataset and enhance their ability to build accurate predictive models.

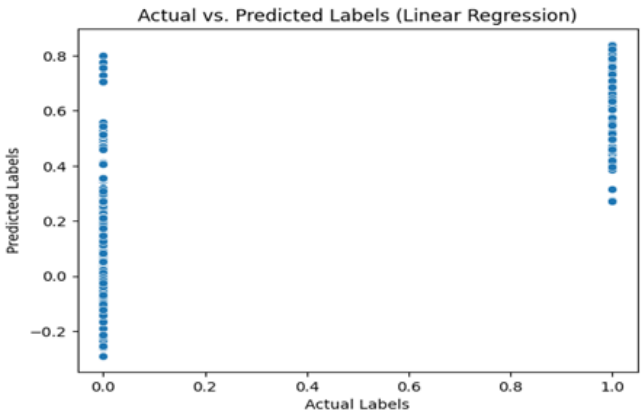


Figure 3 Visualization of Testing Result for Linear Regression

In the above graph, this plot provides a visual representation of how well the Linear Regression model can accurately predict. In this research, the actual values and predicted values are linearly related and have a good fit, with data points on the scatter plot tending to fall along the diagonal line representing perfect predictions.

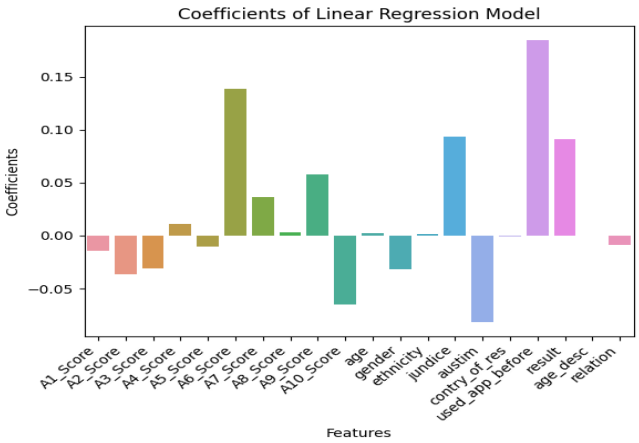


Figure 4 Coefficient of regression

The visualization in Figure 5 is useful for understanding the extent to which each feature influences the target variable in a linear regression model. If there is a feature with a coefficient that is much larger than the others, it indicates that the feature has a more significant impact on the target variable.

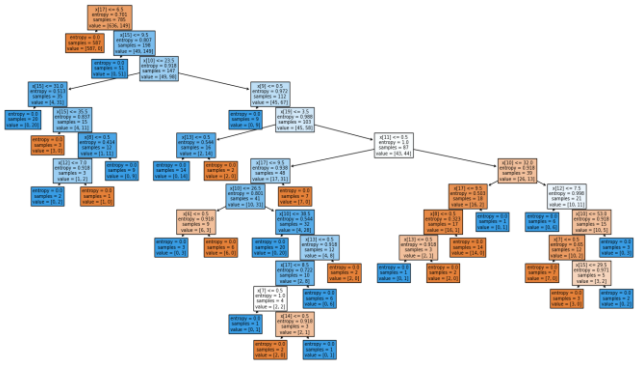


Figure 5 Decision Tree of our proposed C4. 5 model

The generated visualization depicts the structure of a decision tree, including the decision rules at each node and the splits based on different features. Each node in the tree represents a decision based on a specific feature, with branches leading to subsequent nodes or leaf nodes representing prediction outcomes or classes. The filled color in the nodes can provide additional information about the majority class or class distribution within each node.

B. Effect of Data Splitting

The accuracy results for various Split Data scenarios are presented for validation From the table, it can be observed that the accuracy of the C4.5 algorithm is superior to that of the Linear Regression algorithm in making predictions. This conclusion is supported by the accuracy values obtained from different Split Data scenarios. Therefore, it can be inferred that in this study, the C4.5 algorithm outperforms Linear Regression based on the accuracy results provided in the Table 3.

Table 3. Result Comparison

Model	Splitting Scenario		Accuracy (%)	MAE (%)	RMSE (%)
	Train (%)	Test (%)			
LR	80	20	43	24	31
C4. 5	80	20	96	-	-
LR	70	30	44	23	30
C4. 5	70	30	94	-	-
LR	60	40	41	21	28
C4. 5	60	40	97	-	-

VII. Conclusion

Machine learning-based prediction methods, particularly utilizing the C4.5 algorithm and Linear Regression, offer a promising approach to assist in the early diagnosis of autism. These methods have the potential to improve automated detection capabilities, allowing for the prompt provision of appropriate interventions and treatments to individuals with autism, thereby enhancing their quality of life and developmental potential.

Additionally, the identification of factors causing ASD has become more frequent in the era of advanced information technology, further emphasizing the relevance and importance of utilizing machine learning-based prediction methods for early autism diagnosis. Therefore, the study underscores the significance of machine learning techniques in addressing the challenges of early autism detection and intervention.

Further research in the field of early autism disorder prediction using machine learning could explore the integration of various machine learning algorithms to enhance the accuracy and efficiency of predictive models. Additionally, investigating the utilization of different datasets and the impact of diverse demographic and environmental factors on the predictive capabilities of the models could provide valuable insights. Furthermore, exploring the potential of deep learning techniques, such as neural networks, in early autism prediction could be a promising avenue for future research. This could involve examining the performance of deep learning models in comparison to traditional machine learning algorithms, such as the C4.5 algorithm and Linear Regression, to determine their effectiveness in early autism detection. Moreover, conducting longitudinal studies to assess the long-term predictive ability of machine learning models for autism spectrum disorder could provide valuable information on the stability and reliability of these predictive tools over time. Additionally, investigating the feasibility of integrating non-invasive neuroimaging data, such as functional magnetic resonance imaging (fMRI) or electroencephalography (EEG), with machine learning algorithms for early autism prediction could open new possibilities for enhancing the accuracy and early detection of autism spectrum disorder.

References

[1] Chowdhury, K., & Iraj, M. A. (2020, November). Predicting autism spectrum disorder using machine learning classifiers. In 2020 International conference on recent trends on electronics, information, communication & technology (RTEICT) (pp. 324-327). IEEE.

[2] Albahri, A. S., et al. "Early automated prediction model for the diagnosis and detection of children with autism spectrum disorders based on effective sociodemographic and family characteristic features." Neural Computing and Applications 35.1 (2023): 921-947.

[3] Gunawan, W., Wiradiputra, R. A., Sari, A. P., Prayama, D., & Nainggolan, E. R. (2023). Prediction of Cross-Platform and Native Apps Technology Opportunities for Beginner Developers Using C 4.5 and Naive Bayes Algorithms. JOIV: International Journal on Informatics Visualization, 7(4), 2145-2153.

[4] Dwiasnati, S., & Devianto, Y. (2019, July). Utilization of Prediction Data for Prospective Decision Customers Insurance Using the Classification Method of C. 45 and Naive Bayes Algorithms. In Journal of Physics: Conference Series (Vol. 1179, No. 1, p. 012023). IOP Publishing.

[5] Hani'ah, M., Abdullah, M. Z., Sabilla, W. I., Akbar, S., & Shafara, D. R. (2023). Google Trends and Technical Indicator based Machine Learning for Stock Market Prediction. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 22(2), 271-284.

[6] Deng, L., Rattadilok, P., & Xiong, R. (2021, August). A Machine Learning-Based Monitoring System for Attention and Stress Detection for Children with Autism Spectrum Disorders. In Proceedings of the 2021 International Conference on Intelligent Medicine and Health (pp. 23-29).

[7] Choi, E. S., Yoo, H. J., Kang, M. S., & Kim, S. A. (2020). Applying artificial intelligence for diagnostic classification of korean autism spectrum disorder. Psychiatry Investigation, 17(11), 1090-1095. https://doi.org/10.30773/pi.2020.0211

[8] Riandari, F., Sihotang, H. T., & Husain, H. (2022). Forecasting the Number of Students in Multiple Linear Regressions. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 21(2), 249-256.

[9] Muhammad, A., & Defit, S. (2022). Analyzing the use of Social Media by Fashion Designers with K-Means and C45. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 21(2), 463-476.

[10] Ramachandran, A. K. and Martin, V. F. J. B. (2022). Adaptive autism behavior prediction using improved binary whale optimization technique. Concurrency and Computation: Practice and Experience, 35(3). https://doi.org/10.1002/cpe.7511

[11] Shih, C., Lu, C., Chen, G., & Chang, C. (2020). Risk prediction for early chronic kidney disease: results from an adult health examination program of 19,270 individuals. International Journal of Environmental Research and Public Health, 17(14), 4973. https://doi.org/10.3390/ijerph17144973

[12] Garbin, C., Marques, N., & Marques, O. (2023). Machine learning for predicting opioid use disorder from healthcare data: a systematic review. Computer Methods and Programs in Biomedicine, 107573.

[13] P. D. Kusuma, Machine Learning Teori, Program, dan Studi Kasus. Yogyakarta: Deepublish, 2020.

[14] Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., ... & Zhang, J. D. (2020). An introduction to machine learning. Clinical pharmacology & therapeutics, 107(4), 871-885.

[15] Avuçlu, E., & Elen, A. (2020). Evaluation of train and test performance of machine learning algorithms and Parkinson diagnosis with statistical measurements. Medical & Biological Engineering & Computing, 58, 2775-2788.

[16] Li, Y., Zhu, Z., Wu, H., Ding, S., & Zhao, Y. (2020). CCAE: cross-field categorical attributes embedding for cancer clinical endpoint prediction. Artificial Intelligence in Medicine, 107, 101915.

[17] Sachan, S., Almaghrabi, F., Yang, J. B., & Xu, D. L. (2021). Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance. Expert Systems with Applications, 185, 115597.

[18] Z. A. Fikriya, M. I.Irawan, and Soetrisno, “Implementasi Extreme Learning Machine untuk Pengenalan Objek Citra Digital,”J. Sains dan Seni ITS, vol. 6, no. 1, pp. A18–A23, 2017.

[19] Zhu, K., & Wu, J. (2021). Residual attention: A simple but effective method for multi-label recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 184-193).

Ridlan Ahmad - Information System, Institut Teknologi dan Bisnis Stikom, Ambon, Indonesia Email: ahmadridlan83@gamil.com

Muhaimin Hasanuddin - Faculty of Computer Science, Mercu Buana University, Jakarta, Indonesia Email: muhaimin.hasanudin@mercubuana.ac.id

Ofelia Cizela da Costa Tavares - Faculty of Engineering and Science, Computer Science Department, Dili Institute of Technology, Dili-Timor Leste Email: offytasyah1803@gmail.com

Daniel Eliazar Latumaerissa - Faculty of Computer Science, Gadjah Mada University, Yogjakarta, Indonesia Email: daniel.bantuguru@gmail.com