

SALARY PREDICTOR WEB APP USING DECISION TREE REGRESSOR

Sunitha Tappari, Assistant Professor, G. Narayanamma Institute of Technology and Science (For Women), Shaikpet, Hyderabad, India

ABSTRACT:

In today's fiercely competitive job market, the ability to accurately forecast salaries holds paramount importance for both employers and employees. This paper delineates the development of an innovative web-based salary prediction application harnessing the prowess of decision tree regression. Leveraging a rich and expansive dataset curated from Kaggle, comprising an array of pertinent employee attributes including age, gender, educational attainment, job title, and tenure, our research endeavors to furnish a robust and precise tool for salary estimation. Through rigorous data preprocessing methodologies meticulously designed to address missing values and encode categorical variables, in tandem with sophisticated model training strategies, our application is meticulously calibrated to ensure the utmost reliability and accuracy in salary prediction. Moreover, our emphasis on intuitive user interface design enhances accessibility and usability, ensuring seamless engagement and empowering stakeholders with actionable insights into salary determination dynamics. This endeavor underscores the transformative potential of data-driven approaches in redefining contemporary human resource management paradigms, offering tangible solutions to the intricate challenges inherent in salary forecasting.

Key-words: Salary Prediction, Decision tree Regression, Data preprocessing.

INTRODUCTION :

In contemporary labor markets, the accurate estimation of salaries holds significant implications for organizational decision-making and individual career planning [1]. Human resource managers face the intricate task of determining an employee's salary expectation, considering a myriad of factors beyond just past performance or interview performance. Demographic, experiential, and market-related variables are pivotal in shaping an employee's salary expectations. While experienced HR professionals collaborate with departmental managers to navigate this decision-making process, arriving at a satisfactory salary recommendation remains challenging. Salary often drives employee resignations, with higher pay incentivizing retention and stagnant wages prompting job changes. Factors like personal traits, education, and experience shape salary expectations, while Data Science and Machine Learning enable accurate salary predictions by leveraging large datasets [2].

The development of automated decision-making systems could significantly assist in this endeavor, offering precise and data-driven insights to inform salary offers. Although many companies possess internal compensation prediction systems, the lack of access to external data poses a challenge for broader analysis and implementation. Leveraging externally available data sources becomes imperative for refining and enhancing predictive systems aimed at optimizing salary recommendations in recruitment processes [3]. Traditional methods of salary prediction often exhibit limitations in capturing the intricate interplay of factors influencing compensation levels, necessitating the adoption of advanced predictive modeling techniques.

Machine learning algorithms offer promising avenues for salary prediction in recruitment processes. The linear regression algorithm, a cornerstone of supervised learning in machine learning, aims to approximate the mapping function for optimal predictions. Its primary objective lies in constructing a robust model that accurately predicts the dependent attribute based on a set of attribute variables [4]. Linear regression models utilize historical data to establish relationships between various factors and salary levels, providing a straightforward approach to estimating compensation [5].

Random Forest (RF) algorithms, on the other hand, leverage ensemble learning techniques, combining multiple decision trees to yield more accurate predictions by reducing overfitting and enhancing generalization [6]. Leveraging their computational efficiency, they excel in addressing big data classification and regression challenges. The standard approach to constructing random forests

involves a recursive algorithm, where at each node, data is partitioned into subsets based on optimization criteria like the Gini coefficient in CART (Classification and Regression Tree) [7].

In supervised learning for classification and regression, decision trees stand out as a widely favored and dependable algorithm due to their popularity and reliability. Decision trees offer intuitive, interpretable models that partition the feature space based on hierarchical decisions, allowing for transparent insights into salary determinants [8]. Visualizing models provides a graphical overview, aiding comprehension of complex problems. Decision trees in data mining offer insights into attribute impacts on decision-making and generalize rules for assigning tuples to specific classes [9]. Each of these regressors, be it linear regression, Random Forest, or decision trees, brings distinct advantages to salary prediction tasks, empowering HR professionals with valuable tools to optimize compensation strategies and enhance talent acquisition processes.

This paper presents the development of a web-based salary prediction application utilizing decision tree regression, aimed at addressing the inherent complexities in salary estimation. Leveraging a comprehensive dataset obtained from Kaggle, encompassing diverse employee attributes including age, gender, education level, job title, and years of experience, the study seeks to offer a robust and user-friendly tool for salary estimation. By bridging existing gaps in salary prediction methodologies, the research contributes to the advancement of data-driven approaches in human resource management. This paper addresses the need for robust salary prediction methodologies by introducing the development of a web-based salary predictor application utilizing decision tree regression. The primary objective is to provide an accurate and efficient tool for salary estimation, thereby facilitating informed decision-making in human resource management. Specifically, the study aims to:

- Develop a user-friendly web application for salary prediction.
- Utilize decision tree regression as the primary modeling technique.
- Demonstrate the effectiveness of data-driven approaches in salary prediction.

METHODOLOGY :

The methodology integrates diverse phases, ensuring comprehensive coverage and robust execution per established best practices from academia and industry.

DATASET ACQUISITION AND DESCRIPTION :

The dataset utilized in this study, procured from Kaggle, offers a comprehensive repository of employee salary information vital for conducting in-depth analyses. Each record within the dataset represents a unique employee, while the columns encapsulate specific attributes crucial for salary prediction modeling.

Herein lies a detailed exposition of each column:

DESCRIPTION OF DATASET COLUMNS :

1. *Age*: This numeric column represents the age of each employee in years, providing a fundamental demographic characteristic crucial for analysis.
2. *Gender*: Categorized as a binary variable, gender denotes the gender identity of each employee, with values indicating either male or female
3. *Education Level*: This categorical attribute delineates the educational attainment of each employee, ranging from high school to advanced degrees such as a Ph.D.
4. *Job Title*: Represented as a categorical variable, job title signifies the occupational role or position of each employee within the organization, encompassing a diverse range of roles such as manager, analyst, engineer, or administrator.
5. *Years of Experience*: Expressed as a numeric variable, years of experience quantifies the tenure of each employee in the workforce, providing insights into their professional expertise and career trajectory.

6. *Salary*: This numeric column reflects the annual salary of each employee in US dollars, encapsulating various factors such as job title, years of experience, and educational level, thereby serving as a key metric for compensation analysis.

DATA PREPROCESSING :

In data processing, not all variables are equally relevant; feature selection, also termed variable or attribute subset selection, identifies the most impactful attributes for machine learning or statistical analysis. By selecting a subset of variables, this approach streamlines training, aids data interpretation, and mitigates overfitting, enhancing data generalization [10]. Dealing with noisy data is essential for maintaining high classification accuracy. Noise, if not properly handled, can infiltrate various parts of the dataset, leading to increased error rates during analysis [11].

Before model development, the dataset was extensively preprocessed to ensure its quality and suitability for study. This requires numerous processes, including:

Handling missing values: Missing values were imputed using appropriate procedures, such as mean or median imputation, to ensure data integrity.

Encoding categorical variables: Categorical data, such as gender and education level, were encoded using one-hot encoding to turn them into a numerical format appropriate for model training.

Scaling numerical features: Numerical characteristics were scaled to a common range so that no feature could dominate the model training process.

MODEL DEVELOPMENT :

The choice of decision tree regression as the major modeling technique was prompted by its inherent interpretability and ability to adequately capture nonlinear interactions within the dataset [5]. Unlike traditional linear regression models, decision trees provide a flexible framework for modeling detailed interactions between predictor variables, making them ideal for wage prediction where such nonlinearities may exist. Decision tree regression, by recursively partitioning the feature space based on attribute values, makes it easier to identify different patterns and decision boundaries, allowing for accurate and interpretable salary predictions. This methodology strongly corresponds with the study's objectives, highlighting the rationale behind the adoption of decision tree regression as the cornerstone of the modeling approach.

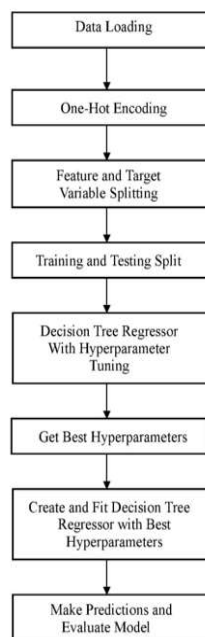


Fig. 1. Decision Tree Regression Model Development

ALGORITHM :

- **Import Libraries:**

The requisite libraries encompassing essential functionalities were imported, including pandas, numpy, and sklearn modules, facilitating seamless data manipulation and model implementation.

- **Read CSV File:**

The dataset, housed within the 'Salary Data.csv' file, was ingested into a pandas DataFrame ('df') utilizing the `pd.read_csv()` function, serving as the foundational data structure for subsequent analyses.

- **One-Hot Encoding:**

Categorical variables such as 'Gender,' 'Education Level,' and 'Job Title' underwent one-hot encoding using the `pd.get_dummies()` function. This transformation facilitated the incorporation of categorical attributes into the modeling process.

- **Split Data:**

The preprocessed dataset was partitioned into distinct feature vectors ('X') and the target variable ('Salary'), thereby delineating the input features and the variable of interest essential for model training.

- **Handle Missing Values in Target Variable:**

Utilizing Simple Imputer from the `sklearn.impute` module, missing values within the target variable ('Salary') were imputed with the mean value, ensuring data integrity and model stability.

- **Split into Training and Testing Sets:**

Employing the `train_test_split()` function from `sklearn.model_selection`, the dataset was stratified into separate training and testing sets, facilitating robust model evaluation and validation.

- **Impute Missing Values in Features:**

Missing values within the feature vectors ('X_train' and 'X_test') were imputed using Simple Imputer, ensuring data completeness and mitigating potential biases in subsequent analyses.

- **Decision Tree Regressor with Hyperparameter Tuning:**

Initialization of a Decision Tree Regressor instantiated the modeling process, followed by the specification of a hyperparameter grid ('param_grid') encompassing parameters such as maximum depth, minimum samples split, and minimum samples per leaf. Grid Search CV, a cross-validation technique, was employed to identify the optimal hyperparameter configuration, thereby enhancing model performance.

- **Get the Best Hyperparameters:**

The grid search results facilitated the retrieval of the optimal hyperparameters, thereby informing subsequent model instantiation and refinement.

- **Create a Decision Tree Regressor with Best Hyperparameters:**

Armed with the optimal hyperparameters, a new instance of the Decision Tree Regressor was instantiated, paving the way for model fitting and subsequent predictive analytics.

- **Fit the Model and Make Predictions:**

The Decision Tree Regressor, endowed with the optimal hyperparameters, was fitted to the training data ('X_train') to imbibe underlying patterns and relationships. Subsequently, predictions about the target variable ('Salary') were generated for the test set, thereby facilitating model evaluation and performance assessment.

- **Evaluate the Model:**

The performance of the trained Decision Tree Regressor was meticulously evaluated, with a comprehensive array of regression metrics including Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-squared computed and analyzed.

WEB APPLICATION DEVELOPMENT :

Following model development, the decision tree regression model was seamlessly integrated into a web-based application using the Streamlit framework.

Streamlit is a powerful and intuitive open-source framework that allows users to create web applications for machine learning and data science projects with ease. It simplifies the process of building interactive web-based tools, enabling developers and data scientists to focus on their core tasks without getting bogged down by the complexities of web development. Streamlit provides an easy and Pythonic way to design and deploy data-driven applications [12], making it accessible to a wide range of users, regardless of their level of expertise in web development.

Streamlit offers a straightforward approach to building interactive web applications with Python, allowing for easy deployment and accessibility. The web application interface enables users to input their demographic and professional details, such as age, gender, education level, job title, and years of experience. Upon submission, the application provides real-time salary predictions based on the deployed decision tree regression model.

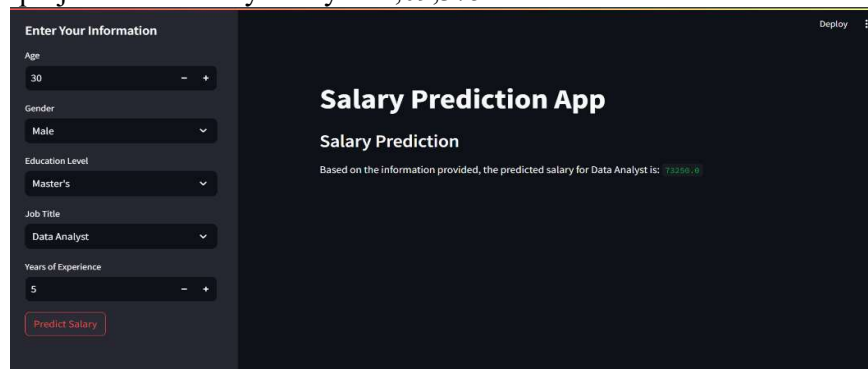
RESULTS :



The screenshot shows a web application titled "Salary Prediction App". On the left, there is a form titled "Enter Your Information" with the following fields: Age (30), Gender (Male), Education Level (Master's), Job Title (Sales Manager), and Years of Experience (7). A "Predict Salary" button is at the bottom of the form. On the right, the app displays "Salary Prediction" and states: "Based on the information provided, the predicted salary for Sales Manager is: 1,09,375.0".

Fig. 2. Streamlit Use Case 1

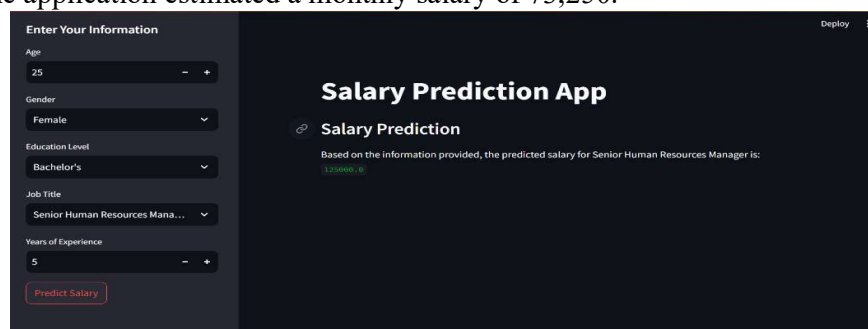
For a user with a master's degree working as a sales manager and possessing 7 years of experience, the application projected a monthly salary of 1,09,375.



The screenshot shows the same "Salary Prediction App" interface. The form on the left has: Age (30), Gender (Male), Education Level (Master's), Job Title (Data Analyst), and Years of Experience (5). The prediction on the right is: "Based on the information provided, the predicted salary for Data Analyst is: 73,250.0".

Fig. 3. Streamlit Use Case 2

When considering a user with a master's degree employed as a data analyst with 5 years of experience, the application estimated a monthly salary of 73,250.



The screenshot shows the "Salary Prediction App" interface. The form on the left has: Age (25), Gender (Female), Education Level (Bachelor's), Job Title (Senior Human Resources Manager), and Years of Experience (5). The prediction on the right is: "Based on the information provided, the predicted salary for Senior Human Resources Manager is: 12,000.0".

Fig. 4. Streamlit Use Case 3

In the case of a user holding a bachelor's degree and serving as a senior HR manager with 5 years of experience, the predicted monthly salary amounted to 1,25,000.

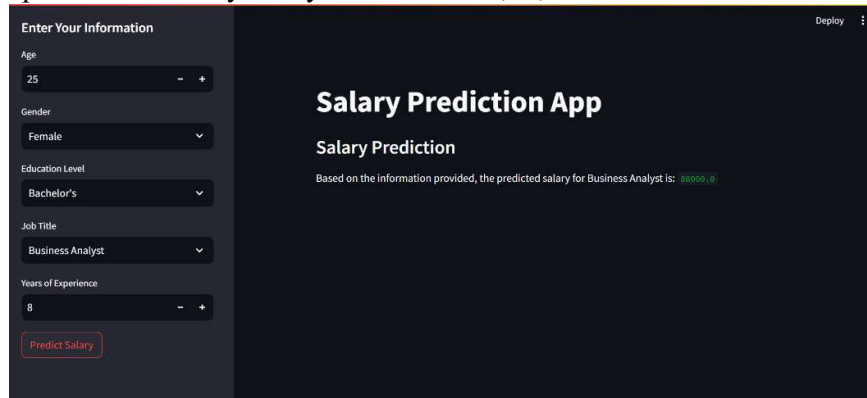


Fig. 5. Streamlit Use Case 4

For a user possessing a bachelor's degree and working as a business analyst with 8 years of experience, the application forecasted a monthly salary of 88,000.

ANALYSIS :

To determine the most appropriate regression technique for analyzing salary data, an experiment was conducted to compare the performance of three regression models, Linear Regression, Random Forest, and decision tree. These regression techniques were evaluated for their ability to predict salary outcomes accurately.

Parameters	Linear Regressor	Random Forest Regressor	Decision Tree Regressor
Mean Absolute Error	1.4871306398374518e+16	11161.31242404899	9723.809949359545
Mean Squared Error	2.473890226805279e+33	309430375.2948222	183021490.9145206
Root Mean Squared Error	4.973821696447591e+16	17590.633169241584	13528.543562206562
R-Squared	-1.0655231887193921e+24	0.8722680237347226	0.9244492506230031

Fig. 6. Comparison of 3 models

The Decision Tree Regressor emerged as the most effective model, demonstrating a remarkable R-squared value nearing 1, indicating a strong correlation between predictor variables and the target variable. Its superior performance was evident in error metrics, showing significant reductions compared to linear regression and random forest, enhancing predictive accuracy. With robustness in capturing underlying patterns, the Decision Tree Regressor surpassed linear regression and the model's performance improvement is attributed to its ability to capture complex relationships through recursive partitioning and fine-tuned hyperparameters. Overall, the Decision Tree Regressor offers a compelling alternative, showcasing low error values and reliability for practical applications, standing out as the optimal choice among regression models.

Following are residual plots for the Linear Regressor, Random Forest Regressor, and Decision Tree Regressor models.

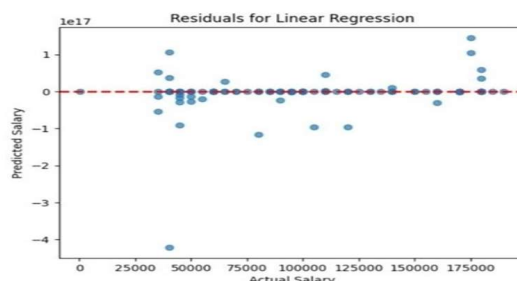


Fig. 7. Residual Plot for Linear Regression

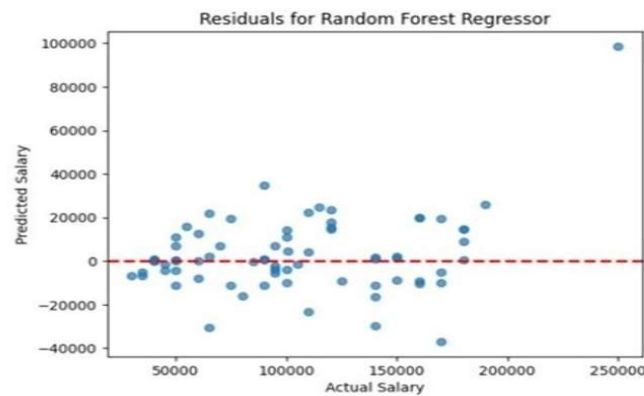


Fig. 8. Residual Plot for Random Forest Regression

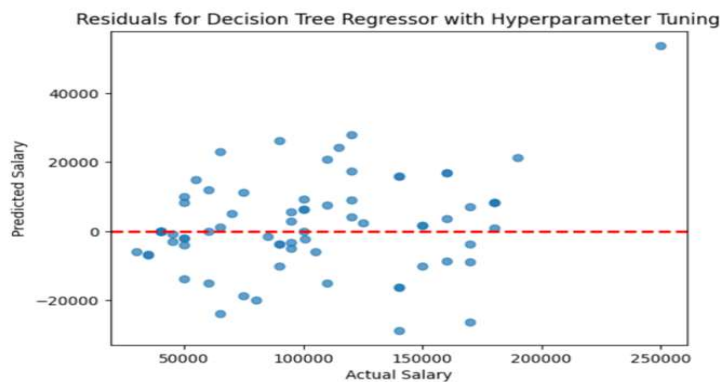


Fig. 9. Residual Plot for Decision Tree Regression

The analysis of residual plots provides valuable insights into the effectiveness of the Decision Tree Regressor in capturing non-linear relationships within the data. Examination of the residual plot reveals a scattered distribution of residuals around the zero line, indicative of the model's capability to effectively capture underlying patterns. This dispersion highlights the model's adeptness at handling non-linear relationships, as evidenced by the varied distribution of data points across the plot. The presence of scattered residuals underscores the Decision Tree Regressor's flexibility in accommodating complex data structures, further enhancing its suitability for predictive modeling tasks.

CONCLUSION :

The comprehensive analysis of salary prediction models, spanning linear regression, the Random Forest Regressor, and the Decision Tree Regressor, has provided invaluable insights into their capabilities. While linear regression encountered challenges in accurately capturing the intricate nuances of salary prediction, as evidenced by elevated error metrics and a pronounced pattern in the residual plot, both the Random Forest Regressor and the Decision Tree Regressor showcased notable performance improvements. The Decision Tree Regressor particularly stood out, demonstrating superior performance metrics compared to both linear regression and the Random Forest Regressor in several aspects. With fine-tuned hyperparameters, the Decision Tree Regressor exhibited reduced error metrics and a strong fit to the data, highlighted by its impressive R-squared value. Additionally, its capacity to capture complex relationships through recursive partitioning contributed significantly to its competitive edge in salary prediction.

REFERENCES :

1. Ignacio Martín, Andrea Mariello, Roberto Battiti, Jose Alberto Hernández.: Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study. *International Journal of Computational Intelligence Systems*, Vol. 11, pp. 1192–1209 (2018).
2. RehamKablaou, Ayed Salman.: Machine Learning Models for Salary Prediction Dataset using Python. In: *2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pp. 143, IEEE (2022).
3. SwapnajitChakraborti.: A Comparative Study of Performances of Various Classification Algorithms for Predicting Salary Classes of Employees. *International Journal of Computer Science and Information Technologies*, Vol. 5 (2), pp. 1964–1972 (2014).
4. D. M. Lothe, Prakash Tiwari, Nikhil Patil, SanjanaPatil, VishwajeetPatil.: Salary Prediction Using Machine Learning. *International Journal of Advance Scientific Research and Engineering Trends*, Vol. 6, Issue 5, ISSN (Online) 2456-0774, (2021).
5. Hai Wang, FeiHao.: An Efficient Linear Regression Classifier. In *Proceedings of the IEEE International Conference*, IEEE (2012).
6. PornthepKhongchai, PokpongSongmuang.: Random Forest for Salary Prediction System to Improve Students' Motivation. In: *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems*, DOI: 10.1109/SITIS.2016.106, pp. 637-642. IEEE, ISBN 978-1-5090-5698-9/16 (2016).
7. Dmitry A. Devyatkin, A, Oleg G. Grigoriev.: Random Kernel Forests. *Federal Research Center - Computer Science and Control, RAS*, 119333 Moscow, Russia (2022).
8. Jocelyn Verna Siswanto, Laurentia Alyssa Castilani, Natasha HartantiWinata, Nathania Christy Nugraha, Noviyanti T M Sagala.: Salary Classification & Prediction based on Job
9. Field and Location using Ensemble Methods. In: *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, pp. 325–330. IEEE, DOI: 10.1109/ICCoSITE57641.2023.10127828, ISBN 978-1-5090-5698-9/16 (2023).
10. Soham Pathak, Indivar Mishra, AleenaSwetapadma.: An Assessment of Decision Tree based Classification and Regression Algorithms. In *Proceedings of the International Conference on Inventive Computation Technologies*, pp. 92. IEEE (2018).
11. Jitendra Kumar Jaiswal, Rita Samikannu.: Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. In *Proceedings of the 6th World Congress on Computing and Communication Technologies (WCCCT)*, pp. 65. IEEE, (2017).
12. Ahmed Mohamed Ahmed, Ahmet Rizaner, Ali HakanUlusoy.: A Decision Tree Algorithm Combined with Linear Regression for Data Classification. In *Proceedings of the 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, IEEE, (2018).
13. IliyasAbduali, AldiyarIbragimov.: Development of a Model for Predicting the Optimal Value for Selected Dataset Using a Web Application. In: *2022 International Conference on Smart Information Systems and Technologies (SIST)*, DOI: 10.1109/SIST54437.2022.9945803, IEEE, ISBN 978-1-6654-6790-2/22 (2022).