**Journal of Computer Science and Technology Studies**
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

**| RESEARCH ARTICLE**

# Machine Learning Approaches to Salary Prediction in Human Resource Payroll Systems

**Jaya Vardhani Mamidala**, University of Central Missouri, Department of Computer Science
**Varun Bitkuri**, Stratford University ,Software Engineer
**Avinash Attipalli**, University of Bridgeport, Department of Computer Science
**Raghuvaran Kendyala**, University of Illinois at Springfield, Department of Computer Science
**Jagan Kurma** , Christian Brothers University, Computer Information Systems
**Sunil Jacob Enokkaren**, ADP, Solution Architect
**Corresponding author**: Jaya Vardhani Mamidala, **Email**: mvardhini29@gmail.com

**| ABSTRACT**

Salary prediction plays a vital role in human resource (HR) management, enabling organizations to streamline payroll systems, improve decision-making, and ensure fair compensation. Accurate salary forecasting supports workforce planning, budgeting, and employee retention strategies. Traditional payroll systems often rely on static rules and historical records, which may not capture complex relationships between employee attributes and income levels. With the advent of machine learning (ML), predictive models have emerged as powerful tools for addressing these limitations in HR payroll systems. This study proposes an Extreme Gradient Boosting (XGBoost) model for salary prediction using the Adult Income Dataset. The method incorporates feature selection following data pretreatment, which includes managing missing values, eliminating outliers, one-hot encoding, and min–max normalization, in order to maintain the most relevant characteristics. To guarantee accurate model assessment, the dataset is separated into training (80%) and testing (20%) subsets. With an AUC-ROC of 0.93 and 91.16% accuracy, precision, recall, and F1-score all at 88%, the suggested XGBoost model demonstrated high predictive performance. The findings show that the XGBoost model performs noticeably better than more conventional models, such as Naïve Bayes (NB) and Support Vector Machine (SVM), making it a dependable and efficient method for predicting salaries in payroll systems. This study highlights the potential of advanced ML techniques to enhance efficiency and accuracy in HR management.

**| KEYWORDS**

Employee Pay Prediction, Employee's Income, Machine Learning, Salary Prediction, Adult Income Dataset

**| ARTICLE INFORMATION**

**ACCEPTED:** 03 October 2025 **PUBLISHED:** 19 October 2025 **DOI:** 10.32996/jcsts.2025.7.10.52

## I. Introduction

The wage of the employee is now the main reason why an employee leaves an organization. Employees regularly switch employment in order to collect the promised remuneration. To keep the firm from losing money, a model has been put up that states that employees should be paid what they want or anticipate from the business [1][2]. In this competitive environment, every person has higher expectations and aspirations. However, it would not be possible to randomly allocate the amount of remuneration that each person would expect. Rather, a system should be established to assess an employee's capability to earn the expected salary. The exact wage cannot be computed, but it can be estimated based on the right data sets [3].

Payroll administration is a vital company operation since it involves the accurate computation and distribution of employee salaries and compensation within legal frameworks. Salary prediction is thus a relevant innovation which a solution to the desires of a job seeker, hiring manager and a student. This aids individuals to make rightful choices on their careers and also aid organization in determining equitable and workable salaries which are amicable. To an employee, it is a means of understanding what the potential payback might be and subsequently result into professional development [4][5]. The employer selects such insights to offer competitive pay that is within his or her financial capacities. The students also have opportunities to reap the benefits by selecting educational routes that would translate into financial rewarding professions. Proper assessment of payment can be used to provide a more effective and just employment market by matching the hopes of employment seekers with the functionality of employers [6][7].

**Journal of Computer Science and Technology Studies**
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

A time series is a set of values or a sequence that a variable takes on at the same time. People think the stock market is one of the most complicated financial systems since it has many stocks whose values move a lot over time [8][9]. ML comprises a set of algorithms or computational approaches that systematically find and reveal patterns in data, allowing it to manage large amounts of heterogeneous data. ML develops automated algorithms that can learn from data and provide predictions or judgments based on the data. Effective ML apps advise businesses on data-driven, empirically supported HR management techniques to reduce employee attrition [10][11]. To the best of the authors' knowledge, however, nobody has utilized ML approaches to anticipate staff turnover in community mental health organizations by analyzing HR data [12][13].

### *A. Motivation and Contribution*

The motivation for this study stems from the critical need to accurately predict individual income levels, which has significant implications for human resource (HR) management, economic analysis, and policy-making. Income prediction helps organizations identify wage disparities, optimize payroll systems, and make informed decisions regarding talent management and compensation strategies. Additionally, leveraging ML techniques, such as XGBoost, enables the effective handling of large, complex datasets and the discovery of patterns that traditional statistical methods may overlook. This research aims to support more egalitarian and data-driven decision-making processes in organizational and socioeconomic contexts by enhancing forecast accuracy and identifying key income factors. This study adds a number of significant insights to the topic of payroll systems for HR:

- Utilized The UCI ML Repository's Adult Income Dataset, containing 48,842 records and 14 attributes (8 categorical, 6 continuous) across 42 nations, providing a diverse and comprehensive dataset for income prediction
- Performed comprehensive data preparation, such as one-hot encoding, outlier elimination, and managing missing values, and normalization, to ensure high-quality input for modeling.
- Delivered elaborate exploratory data analysis with graphs, with emphasis on the relationships between demographic, work characteristics, and income groups.
- Produced a precise income prediction model with the Extreme Gradient Boosting (XGBoost) algorithm, which outperforms the traditional classifiers.
- A variety of evaluation measures, including as ROC-AUC, F1-score, accuracy, precision, and recall, were used to evaluate the model's performance.

### *B. Justification and Novelty*

This research is justified by the growing need for accurate salary prediction models to support HR payroll systems, workforce planning, and compensation management. Conventional payroll methods are often inefficient and fail to accurately reflect the complex and non-linear interactions among demographic, educational, and occupational variables, resulting in inefficiencies and inaccurate classifications. The novelty here is that the Extreme Gradient Boosting (XGBoost) model has been used, and it uses ensemble learning to outperform the traditional classifiers, like Naive Bayes (NB) and Support Vector Machine (SVM), on predictive performance. This study demonstrates that XGBoost can enhance not only predictive accuracy (91.16%) but also balance precision, recall, and F1-score, providing a more valid and viable solution for real-life payroll management systems.

### *C. Structure of the paper*

The structure of this paper is organized as follows: Section II reviews the summarized literature on ML and the prediction of salaries in payroll systems. Section III describes the methodology, including dataset description, preprocessing, and the proposed model. Section IV discusses the results and comparative analysis with existing models. Finally, Section V concludes the study and provides guidelines for future research.

## II. LITERATURE REVIEW

A thorough review and analysis of key research studies on salary prediction for HR payroll systems was conducted to guide and support the development of this study.

Zuo et al. (2019) examine the supply and demand situation for local talent using relevant information from the city's talent market. Wavelet threshold denoising is used to analyze the data, and an NAR neural network model is created for prediction. To anticipate the employment situation, the GM (1,1) forecasting model is integrated with data such as average income, employment satisfaction, and the employment rate of Chinese students. The number of applications, academic criteria, students admitted, and a city's rapidly expanding industry are all taken into consideration while creating the SOM neural network model [14].

**Journal of Computer Science and Technology Studies**
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

Viroonluecha and Kaewkiriya (2018) build the Salary Predictor System to forecast Thai employees' monthly salaries using the DL technique, which has garnered a lot of interest in the ML area lately. The data was taken from a popular job-search website that has over 1.7 million users. To build and analyze this model, personal data from the first five months of 2018 was used. They evaluated the results against similar algorithms, such as GBT and RF. After comparison, DL was subjected to the feature selection techniques. The ideal R-squared result from combining feature selection and DL was 0.462, and the quick runtime was 15.37 seconds [15].

Dutta, Halder and Dasgupta (2018) developed a novel salary prediction engine using the ADZUNA job postings dataset, which contained more than 240,000 records of job titles, descriptions, locations, and company details. The study implemented DT, RF, KNN, and SVM. After extensive preprocessing, including feature selection and salary normalization, the findings indicated that whereas decision trees had an accuracy of 84.4% with a macro F1-score of 0.615, the random forest model outperformed the others with a weighted F1-score of 0.869, an accuracy of 87.3%, a mean squared error of 329.12, and a mean absolute error of 5.04. The authors concluded that job title, description, and location were the strongest predictors of salary, whereas contract type and time had a negligible impact [16].

Martín et al. (2018) examines 4,000 job offers from an IT employment website in Spain. They conclude that experience is more beneficial than education, then utilize tree-based ensembles to develop an acceptable salary-range classifier and create 5 profile groups based on the necessary competencies. This study included a variety of models, including voting classifiers based on all or some of the following models: LR, NN, MLPs, SVMs, RF, and adaptive boosting. According to experiments, DT-based ensembles perform better overall, and a voting committee using them achieves an accuracy of about 84% [17].

Shankar et al. (2018) Employee attrition is a major problem for businesses, particularly when important personnel, technical staff, and skilled workers depart in search of better opportunities elsewhere. The cost of replacing a skilled worker is incurred. Therefore, analyze the common causes of employee attrition using both historical and current employee data. To prevent employee attrition, well-known classification algorithms were applied to the HR data, including DT, LR, SVM, KNN, RF, and NB approaches [18].

Sisodia, Vishwakarma and Pujahari (2017) order to predict the employee attrition rate, try creating a model using a data about HR analytics from the Kaggle website. In order to illustrate the relationship between characteristics, a heatmap and a correlation matrix are created. The experimental portion generates a histogram that displays the disparity between the departing workers and their departments, pay, satisfaction levels, and other factors. Use five distinct ML methods, including RF, KNN, NB classifier, C 5.0 DT classifier, and linear SVM, for prediction purposes. This essay suggests the factors that maximize employee churn in a particular organization [19].

Khongchai and Songmuang (2016) proposes a technique to increase pupils' incentive to study by forecasting earnings. A prediction model with seven characteristics is created using the decision tree approach. used a 10-fold cross-validation using 13,541 records of graduating student data to assess the system's effectiveness. Overall, 41.39% accuracy was achieved. Additionally, use surveys with a sample of 50 students to assess the efficacy of the system. Based on the results, the strategy can help achieve success in engaging students in gaining knowledge and give them some confidence in their future. The last point of the sample students indicated that they were pleased with the technology as they thought the prediction findings were legible and understandable, and the system was simple to use [20].

Table I summaries the modern findings in the area of salary prediction, specifying the models employed, data sets utilized, key findings, and the difficulties encountered.

TABLE I. COMPARATIVE ANALYSIS STUDIES ON SALARY PREDICTION USING MACHINE LEARNING

| Author | Proposed Work | Results | Key Findings | Limitations / Recommendation |
|---|---|---|---|---|
| Zuo et al. (2019) | Analyzed talent market data of a city utilizing SOM, GM(1,1) model, NAR neural network, and wavelet threshold denoising. | Employment satisfaction, average wage, student employment rate, and demand for talent in the labour market. | Combined NAR, GM(1,1), and SOM models provide effective prediction of employment trends and industry demand. | Limited to data from a single city; recommend extending to multi-city datasets for broader applicability. |
| Viroonluecha | Salary Predictor System using | Best $R^2$ = 0.462 with | Feature selection | $R^2$ value is relatively low; |

**Journal of Computer Science and Technology Studies**
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts



| | | | | |
|---|---|---|---|---|
| and Kaewkiriya (2018) | Deep Learning with feature selection on Thai job dataset (1.7M users). | runtime = 15.37s. | combined with Deep Learning improved accuracy and efficiency. | model needs improvement in predictive power. Can extend to larger/varied datasets. |
| Dutta, Halder and Dasgupta (2018) | Salary prediction engine using ADZUNA dataset (240K+ job postings). Models: Decision Trees, RF, KNN, SVM. | RF achieved accuracy = 87.3%, MAE = 5.04, MSE = 329.12, Weighted F1 = 0.869. | Job title, description, and location were strongest predictors. Contract type/time had little effect. | Limited to ADZUNA dataset; cross-platform generalization needed. Explore deep learning for better performance. |
| Martín et al. (2018) | Salary-range classifier on 4,000 Spanish IT job offers using multiple ML models (LR, KNN, MLP, SVM, RF, AdaBoost, Voting). | Tree-based ensembles ≈ 84% accuracy; Voting classifier best. | Experience more rewarded than education; identified five skill-based clusters. | Dataset size small (4,000 jobs). Larger dataset and cross-domain testing recommended. |
| Shankar et al. (2018) | Employee attrition prediction using HR data and ML models (DT, LR, SVM, KNN, RF, NB). | Models applied, results highlight attrition patterns (metrics not specified). | Attrition causes identified; ML useful in HR analytics. | Quantitative performance not deeply reported. Future work: deeper feature engineering and ensemble methods. |
| Sisodia, Vishwakarma and Pujahari (2017) | Employee churn prediction using Kaggle HR dataset with correlation, heatmap, and ML models (SVM, C5.0 DT, RF, KNN, NB). | Predictions generated (accuracy metrics not specified). | Correlation insights: churn relates to salary, department, satisfaction level, etc. | More detailed evaluation metrics needed. Recommendation: apply advanced DL methods and real-world datasets. |
| Khongchai and Songmuang (2016) | Salary prediction for student motivation using a Decision Tree on graduate data (13,541 records). | Accuracy = 41.39% (10-fold CV). Student survey: positive satisfaction. | System boosts motivation and provides simple, interpretable results. | Predictive accuracy low. Recommendation: adopt advanced ML/DL models and richer features. |

## III. Research Methodology

The study's methodology began with data collection using the Adult Income Dataset. The data processing involved handling missing data, removing outliers, applying one-hot encoding, and performing min-max normalization. Following this, feature selection was employed to retain the most useful attributes and eliminate redundancy, thereby enhancing model accuracy and efficiency. To guarantee accurate model evaluation, the data were then divided into two groups: training (80%) and test (20%). Lastly, the suggested Extreme Gradient Boosting (XGBoost) algorithm was created and trained to produce reliable results for salary prediction. The main criteria utilized to evaluate how successfully the HR payroll system predicted wages were the accuracy, precision, recall, F1-score, and ROC curves. Figure 1 shows the full procedure.

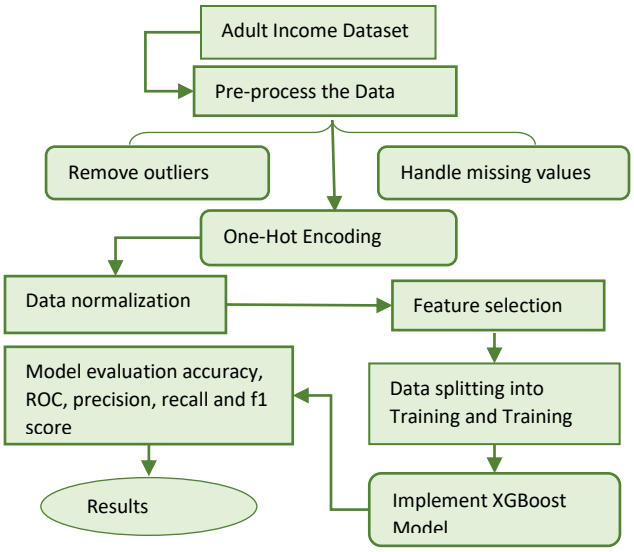


Fig. 1. Proposed flowchart for Salary Prediction

The next section goes into great detail on the suggested flowchart for predicting wages in HR payroll systems.

### *A. Data Collection*

This study utilized the Kaggle dataset on adult income. Based on characteristics including age, education, occupation, and weekly hours worked, it is frequently used in classification tasks to determine if an individual makes more than $50,000 per year (>50K) or less than that much (<=50K). For 42 countries, the data collection has 14 characteristics and 48,842 distinct records. Age, education, nationality, marital status, relationship status, occupation, job classification, gender, race, weekly working hours, capital loss, and capital gain are among the 14 characteristics. These are composed of 8 categorical and 6 continuous components. Data visualizations such as bar plots and heatmaps were used to examine distribution, feature correlations, etc., and are given below:

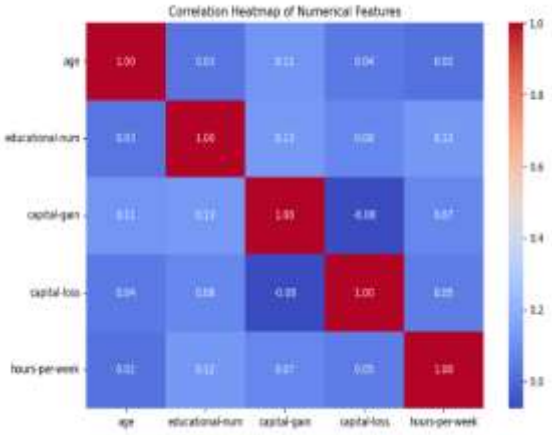

Fig. 2. Heatmap Label's Pearson Correlation Coefficients

Figure 2 shows a correlation heatmap displaying the relationships between 5 numerical features: Age, weekly hours worked, level of education, capital gain, and capital loss. A high positive correlation (closer to 1.00) is indicated by red in the chart, whereas a negative or weak correlation (near to 0.00) is indicated by blue. As anticipated, the diagonal displays a perfect correlation of 1.00 between each variable and itself. The off-diagonal values show that most of the features have very weak correlations with each other, with all values below 0.13. This suggests that none of these variables has a strong linear relationship.

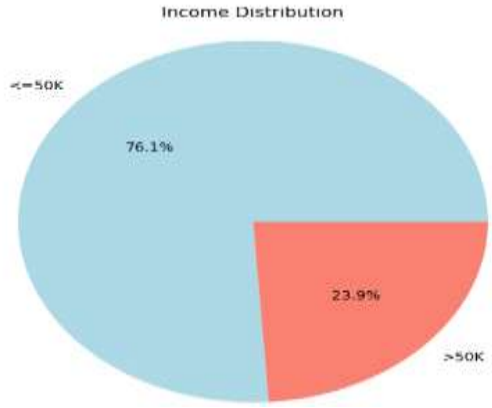


Fig. 3. Pie chart of Income Distribution

Figure 3 shows the pie chart of income distribution, illustrating the proportion of individuals in a dataset based on two income brackets: less than or equal to $50K and greater than $50K. The chart indicates that the majority of individuals, 76.1%, fall into the lower income bracket ($≤ $ 50K), while a smaller segment, 23.9%, belongs to the higher income bracket (>$50K).

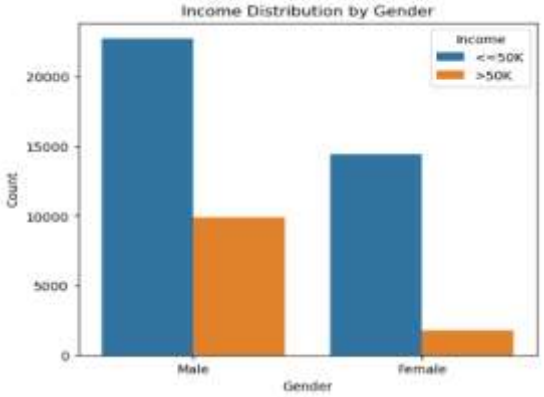


Fig. 4. Income Distribution by Gender

Figure 4 bar chart shows Income Distribution by Gender, compares the number of individuals across two income brackets, less more than $50K and less than or equal to $50K, for both males and females. The chart indicates that males outnumber females in both income categories. Specifically, the proportion of men who make over $50,000 is far more than the number of females in the same income bracket. Conversely, a larger number of females are in the $50K or less income bracket, even though the dataset's overall female population seems to be lower than its overall male population.

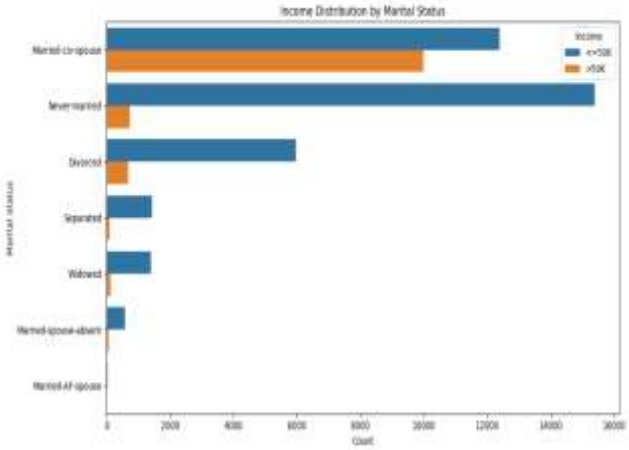


Fig. 5. Income Distribution by Marital Status

Income Distribution by Marital Status Figure 5, the chart shows that Never-married individuals are predominantly in the ≤50K bracket, while Married-civ-spouse individuals are more evenly distributed, with a higher proportion in the >50K bracket. Other marital statuses have fewer individuals, mostly in the ≤50K category.

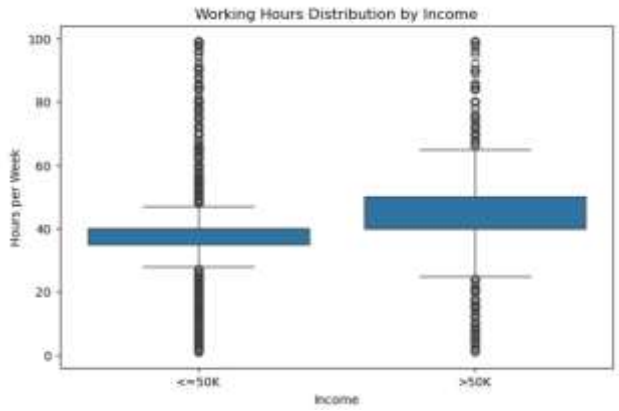


Fig. 6. Working Hours Distribution by Income

Figure 6 working Hours Distribution by Income. The median working hours for the >$50K collection is higher compared to the ≤$50K group, with a larger interquartile range, indicating that higher-income individuals generally work longer hours with greater variability. Both groups exhibit a substantial number of outliers beyond the whiskers, representing individuals with extremely short or long working hours.

### *B. Data Pre-Processing*

The Adult Income Dataset was collected, concatenated, and cleansed before extracting the relevant features. During preprocessing, missing values and outliers were removed, followed by data transformation and normalization. The main preprocessing steps are as follows:

- **Handle missing value:** The dataset has been subjected to several algorithmic adjustments in order to address the missing values for categorical variables, labor class, occupation, and native country.
- **Remove Outliers:** Outliers were handled to ensure data integrity and consistency, preparing the dataset for further investigation and model training.

### *C. One-Hot Encoding for Data Encoding*

Converting data from one format to another, generally with the aim of increasing efficiency, is known as data encoding, which ensures compatibility and facilitates transmission and storage. A data preparation method used in ML called "one-hot encoding" transforms categorical data into a computer-understandable numerical format.

### D. Data Normalization using Min-Max

The data was normalized using the min-max technique, which maintained the values between 0 and 1. This was done to lessen the effects of outliers and make the classifiers that were used work better. Used the mathematical approach Equation (1) to do the normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \cdots$$

Where X stands for the feature's initial value, $X'$ for its normalised value, $X_{min}$ for its minimum value, and $X_{max}$ for its maximum value.

### E. Feature Selection

Feature selection is to constructing an ML model, the process of identifying and selecting the most relevant features (variables) from a dataset. Its main objectives are to decrease overfitting and increase model accuracy, minimize computational costs, and streamline models by eliminating unnecessary, noisy, or redundant features.

### F. Data Splitting

The process of breaking up a dataset into more manageable, discrete subsets is known as data splitting, and it is usually used to train, validate, and test ML models. There are training and testing sets inside the dataset. 80% of the data is made accessible for training, while 20% is used for testing.

### G. Proposed Extreme Gradient Boosting (XGBoost) Model

XGBoost is an ensemble learning algorithm that is utilized to predict using DT [21]. Regression problems can be solved by minimizing a loss function that determines the discrepancy between the target's actual and expected values. The XGBoost regression mathematical model can be defined as Equation (2):

$$y = f(x) \quad \cdots$$

Where y is the forecasted property price, x is an input feature (squares of the house, number of bedrooms, etc.), and f(x) defines the XGBoost model that predicts y, using x as input features. XGBoost trains a number of trees in an ensemble to minimize the mean squared error (MSE) loss function in order to compute f(x). The model aggregates the forecasts of different DT to get a final forecast. The overall model of the XGBoost regression can be represented as Equation (3):

$$y = \sum(k = 1\ to\ K)\ fk(x) \quad \cdots$$

$x$ is a predictor variable, $x$ denotes a specific variable to be determined, while $f_k(x)$ denotes the prediction of the $K$-th tree of the decision trees within the ensemble, and $K$ is the number of decision trees in the ensemble. Each tree is predicted as a weighted average of the leaf values of the tree, learned during training. The prediction of the XGBoost model for a given input $x$ is obtained by summing the predictions of all the DT in the ensemble.

### H. Evaluation Metrics

The efficacy of the proposed paradigm was evaluated using various performance metrics. By contrasting the actual values with the anticipated outputs of the trained models, True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) were determined. Based on these measures, the evaluation matrix, which includes F1-score, recall, accuracy, and precision, is displayed below:

*1) Accuracy*

The number of instances that the trained model accurately predicted relative to the total quantity of instances in the dataset (input samples) is expressed as Equation (4)-

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+Fp+TN+FN}} \quad \cdots$$

*2) Precision*

Precision is the percentage of positive instances successfully predicted out of all positive instances predicted by the model. Precision indicates. How good the classifier is in predicting the positive classes is expressed as Equation (5)-

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad \cdots$$

*3) Recall*

This measure is the rate of positive occurrences that were accurately predicted for all cases that ought to have turned out positively. It is expressed mathematically as Equation (6)-

$$Recall = \frac{\mathrm{TP}}{TP+FN} \quad \cdots$$

*4) F1 score*

It combines the precision and recalls harmonic means, which means, it helps to balance recall and precision. Its range is [0, 1]. Mathematically, it is given as Equation (7)-

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \cdots$$

*5) Receiver Operating Characteristic Curve (ROC)*

The percentage of cases that are correctly labelled as positive against those that are wrongly classed as positive for numerous decision cut-off points is shown in a visual representation known as the ROC. FPR is equivalent to 1-specificity, but TPR is sometimes known as sensitivity or recall.

## IV. RESULTS AND DISCUSSION

The models and tests were implemented in Python 3.7 and run on the Microsoft Windows 10 Pro Operating System (OS). used GPU graphics driver version 431.60 and CUDA version 10.1. The TensorFlow (v2.0) and Keras (v2.3) Python packages were utilized, along with scikit-learn, SciPy, and Matplotlib libraries. The proposed model was trained using Table II provides a summary of the Adult Income Dataset and the major performance indicators, F1-score, recall, accuracy, and precision, for example. Using precision, recall, and F1-score all at 88% and an accuracy of 91.16%, the proposed XGBoost model showed balanced classification performance. The AUC-ROC score of 0.93 further demonstrates its strong discriminative ability between income groups. Overall, the results confirm XGBoost as a robust and reliable model for salary prediction in payroll systems.

TABLE II. EXPERIMENT RESULTS OF PROPOSED MODELS FOR SALARY PREDICTION ON ADULT INCOME DATASET

| Performance Matrix | Extreme Gradient Boosting (XGBoost) Model |
|---|---|
| Accuracy | 91.16 |
| Precision | 88 |
| Recall | 88 |
| F1-score | 88 |
| AUC-ROC | 0.93 |

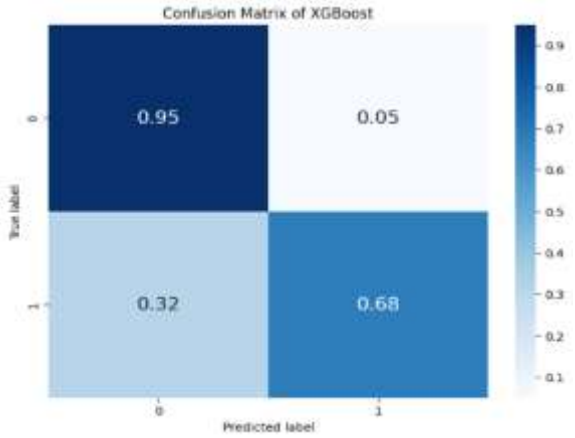


Fig. 7. Confusion matrix for XGBoost Model

The confusion matrix in Figure 7 shows how well the classification model performed on the test data. The model exhibits high predictive power for class 0, with 95% of instances correctly classified and only 5% misclassified as class 1. In contrast, for class 1, the model correctly identifies 68% of the cases, but 32% are misclassified as class 0, indicating a higher error rate for this class. Overall, the model demonstrates excellent accuracy in detecting class 0 while showing relatively weaker performance for class 1, suggesting some imbalance in predictive capability between the two classes.

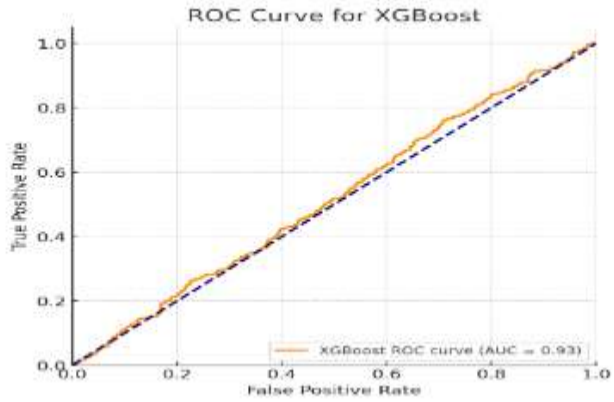


Fig. 8. Precision-Recall analysis of the XGBoost Model

The ROC curve for the XGBoost model is displayed in Figure 8, demonstrating its capacity to differentiate between classes over a range of threshold values. The curve lies well above the diagonal reference line, which represents random classification, indicating strong discriminatory power. The model performed quite well in differentiating between positive and negative pay classes, as evidenced by its AUC value of 0.93. This high AUC value suggests that for salary prediction, the proposed XGBoost model is incredibly reliable and successful. with the least amount of overlap between the TP and FPR.

### *A. Comparative analysis*

In order to support the suggested XGBoost model's efficacy, the comparative accuracy analysis is provided in Table III with other existing models. Although the Adult Income Dataset is split into four distinct groups that differ in various aspects, the accuracy comparison of various predictive models used in an HR payroll system to predict salary across the Adult Income Dataset demonstrates that there are significant differences in the performance of the dataset's prediction models. Naive Bayes (NB) classifier could reach a relatively low accuracy of 78 percent in prediction. Support Vector machine (SVM) produced a better result of 85.52, which demonstrates the fact that it can be used successfully with complex features. The proposed XGBoost model outperformed the other two and demonstrated the highest accuracy at 91.16, which suggests it possesses a more favorable ability to learn and successfully understand the general nonlinear trends in the dataset. The analysis presented in this comparison clearly shows that XGBoost is the most suitable model among the methods analyzed for predicting salaries.

TABLE III. ACCURACY COMPARISON OF DIFFERENT PREDICTIVE MODELS OF SALARY PREDICTION FOR HUMAN RESOURCE PAYROLL SYSTEMS USING THE ADULT INCOME DATASET

| Models | Accuracy |
|---|---|
| NB[22] | 78 |
| SVM[23] | 85.52 |
| XGBoost | 91.16 |

The accuracy of the proposed XGBoost model is 91.16 percent, which outperforms the performance of other traditional models. This superb accuracy indicates the model's high capacity to successfully predict salary categories, which qualifies it as a more legitimate and effective choice in HR payroll methods. There is also increased accuracy, which means that payroll management decision-making is enhanced. It also reduces the chances of being misclassified compared with other models. Overall, it indicates that XGBoost is the most efficient and trustworthy model for salary predictions.

## *V. CONCLUSION AND FUTURE STUDY*

The value of employee pay prediction in HR management lies in its ability to facilitate reasonable compensation, effective payroll management, and informed workforce planning. An Extreme Gradient Boosting (XGBoost) regression model was proposed in this paper to model salary using the Adult Income Dataset, which was preprocessed, features were selected, and the model was trained. The model was 91.16% accurate, 88% precise, recall and F1-score with an AUC-ROC of 0.93. These results confirm its power and equal predictive ability in classifying salary levels. The XGBoost, as compared to other standard models such as NB and SVM, was noted to be a better model because it exhibited a better performance in detecting complex nonlinear data characteristics. The model, however, was found to be more predictive in the category of ≤ $50K incomes rather than the category of > $50K, which was due to a lack of class balance within the dataset. Despite this weakness, the approach provides beneficial insights to organizations, enabling more accurate, fair, and data-driven decisions in payroll systems and workforce

analytics. Overall, this paper demonstrates that XGBoost can be a credible and useful model with the potential to positively impact predictive performance, reduce misclassification, and enhance compensation management in a real-life HR environment.

Subsequent research will enhance the study's quality by employing resampling or cost-sensitive methods, incorporating more socio-economic data, and utilizing sophisticated feature engineering with hyperparameter optimization to make the research more scalable and applicable to the real world.

**References**


[1] I. Martín, A. Mariello, R. Battiti, and J. A. Hernández, "Salary prediction in the IT job market with few high-dimensional samples: A Spanish case study," *Int. J. Comput. Intell. Syst.*, 2018, doi: 10.2991/ijcis.11.1.90.

[2] S. S. S. Neeli, "Serverless Databases : A Cost-Effective and Scalable Solution," *IJIRMPS*, vol. 7, no. 6, 2019.

[3] P. Khongchai and P. Songmuang, "Implement of salary prediction system to improve student motivation using data mining technique," in *Proceedings - 11th 2016 International Conference on Knowledge, Information and Creativity Support Systems, KICSS 2016*, 2017. doi: 10.1109/KICSS.2016.7951419.

[4] A. Tambde and D. Motwani, "Employee churn rate prediction and performance using machine learning," *Int. J. Recent Technol. Eng.*, 2019, doi: 10.35940/ijrte.B1134.0982S1119.

[5] S. S. S. Neeli, "The Significance of NoSQL Databases: Strategic Business Approaches and Management Techniques," *J. Adv. Dev. Res.*, vol. 10, no. 1, 2019.

[6] X. Chen, L. Zhang, Y. Liu, and K. Kenthapadi, "How LinkedIn economic graph bonds information and product: Applications in LinkedIn salary," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018. doi: 10.1145/3219819.3219921.

[7] S. Chakraborti, "A Comparative Study of Performances of Various Classification Algorithms for Predicting Salary Classes of Employees," *Int. J. Comput. Sci. Inf. Technol.*, 2014.

[8] E. Rombaut and M.-A. Guerry, "Predicting voluntary turnover through human resources database analysis," *Manag. Res. Rev.*, vol. 41, no. 1, pp. 96–112, Jan. 2018, doi: 10.1108/MRR-04-2017-0098.

[9] A. Kushwaha, P. Pathak, and S. Gupta, "Review of optimize load balancing algorithms in cloud," *Int. J. Distrib. Cloud Comput.*, vol. 4, no. 2, pp. 1–9, 2016.

[10] Gopi, "Zero Trust Security Architectures for Large-Scale Cloud Workloads," *Int. J. Res. Anal. Rev.*, vol. 5, no. 2, pp. 960–965, 2018.

[11] A. Thapliyal, P. S. Bhagavathi, T. Arunan, and D. D. Rao, "Realizing Zones Using UPnP," in *2009 6th IEEE Consumer Communications and Networking Conference*, IEEE, Jan. 2009, pp. 1–5. doi: 10.1109/CCNC.2009.4784867.

[12] S. N. Mishra, D. R. Lama, and Y. Pal, "Human Resource Predictive Analytics (HRPA) For HR Management In Organizations," *Int. J. Sci. Technol. Res.*, 2016.

[13] A. M. Esmaieeli Sikaroudi, R. Ghousi, and A. Sikaroudi, "A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)," *J. Ind. Syst. Eng.*, vol. 8, no. 4, pp. 106–121, 2015.

[14] J. Zuo, C. Zhang, J. Chen, Y. Wu, Z. Liu, and Z. Li, "Artificial Intelligence Prediction and Decision Evaluation Model Based on Deep Learning," in *2019 International Conference on Electronic Engineering and Informatics (EEI)*, IEEE, Nov. 2019, pp. 444–448. doi: 10.1109/EEI48997.2019.00102.

[15] P. Viroonluecha and T. Kaewkiriya, "Salary Predictor System for Thailand Labour Workforce using Deep Learning," in *ISCIT 2018 - 18th International Symposium on Communication and Information Technology*, 2018. doi: 10.1109/ISCIT.2018.8587998.

[16] S. Dutta, A. Halder, and K. Dasgupta, "Design of a novel prediction engine for predicting suitable salary for a job," *Proc. - 2018 4th IEEE Int. Conf. Res. Comput. Intell. Commun. Networks, ICRCICN 2018*, pp. 275–279, 2018, doi: 10.1109/ICRCICN.2018.8718711.

[17] I. Martín, A. Mariello, R. Battiti, and J. A. Hernández, "Salary prediction in the IT job market with few high-dimensional samples: A Spanish case study," *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, pp. 1192–1209, 2018, doi: 10.2991/ijcis.11.1.90.

[18] R. S. Shankar, J. Rajanikanth, V. V. Sivaramaraju, and K. V. S. S. R. Murthy, "Prediction of Employee Attrition Using Datamining," in *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, IEEE, Jul. 2018, pp. 1–8. doi: 10.1109/ICSCAN.2018.8541242.

[19] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, IEEE, Nov. 2017, pp. 1016–1020. doi: 10.1109/ICICI.2017.8365293.

[20] P. Khongchai and P. Songmuang, "Improving students' motivation to study using salary prediction system," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, IEEE, Jul. 2016, pp. 1–6. doi: 10.1109/JCSSE.2016.7748896.

[21] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A Comparative Analysis of XGBoost," *Artif. Intell. Rev.*, Nov. 2019.

[22] S. M. Bekena, "Using decision tree classifier to predict income levels," *Econ. Policy*, no. 2116, pp. 0–33, 2017.

*[23]* M. Topiwalla, "Machine Learning on UCI Adult data Set Using Various Classifier Algorithms and Scaling Up the Accuracy Using Extreme Gradient Boosting," University of SP Jain School of Global Management, 2017.*Pabbineedi, S., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Tyagadurgam, M. S. V., & Gangineni, V. N. (2023). Scalable Deep Learning Algorithms with Big Data for Predictive Maintenance in Industrial IoT. International Journal of AI, BigData, Computational and Management Studies, 4(1), 88-97.*

*[24] Chalasani, R., Vangala, S. R., Polam, R. M., Kamarthapu, B., Penmetsa, M., & Bhumireddy, J. R. (2023). Detecting Network Intrusions Using Big Data-Driven Artificial Intelligence Techniques in Cybersecurity. International Journal of AI, BigData, Computational and Management Studies, 4(3), 50-60.*

*[25] Vangala, S. R., Polam, R. M., Kamarthapu, B., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2023). A Review of Machine Learning Techniques for Financial Stress Testing: Emerging Trends, Tools, and Challenges. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 4(1), 40-50.*

*[26]Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Tyagadurgam, M. S. V., Gangineni, V. N., & Pabbineedi, S. (2023). A Survey on Regulatory Compliance and AI-Based Risk Management in Financial Services. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 4(4), 46-53.*

*[27]Bhumireddy, J. R., Chalasani, R., Vangala, S. R., Kamarthapu, B., Polam, R. M., & Penmetsa, M. (2023). Predictive Machine Learning Models for Financial Fraud Detection Leveraging Big Data Analysis. International Journal of Emerging Trends in Computer Science and Information Technology, 4(1), 34-43.*

*[28]Gangineni, V. N., Pabbineedi, S., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Tyagadurgam, M. S. V. (2023). AI-Enabled Big Data Analytics for Climate Change Prediction and Environmental Monitoring. International Journal of Emerging Trends in Computer Science and Information Technology, 4(3), 71-79.*

*[29]Polam, R. M. (2023). Predictive Machine Learning Strategies and Clinical Diagnosis for Prognosis in Healthcare: Insights from MIMIC-III Dataset. Available at SSRN 5495028.*

*[30]Narra, B., Gupta, A., Polu, A. R., Vattikonda, N., Buddula, D. V. K. R., & Patchipulusu, H. (2023). Predictive Analytics in E-Commerce: Effective Business Analysis through Machine Learning. Available at SSRN 5315532.*

*[31]Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Polu, A. R., Vattikonda, N., & Gupta, A. K. (2023). Advanced Edge Computing Frameworks for Optimizing Data Processing and Latency in IoT Networks. JOETSR-Journal of Emerging Trends in Scientific Research, 1(1).*

*[32]Patchipulusu, H. H. S., Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., & Buddula, D. V. K. R. (2023). Opportunities and Limitations of Using Artificial Intelligence to Personalize E-Learning Platforms. International Journal of AI, BigData, Computational and Management Studies, 4(1), 128-136.*

*[33]Madhura, R., Krishnappa, K. H., Shashidhar, R., Shwetha, G., Yashaswini, K. P., & Sandya, G. R. (2023, December). UVM Methodology for ARINC 429 Transceiver in Loop Back Mode. In 2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC) (pp. 1-7). IEEE.*

*[34]Shashidhar, R., Kadakol, P., Sreeniketh, D., Patil, P., Krishnappa, K. H., & Madhura, R. (2023, November). EEG data analysis for stress detection using k-nearest neighbor. In 2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS) (pp. 1-7). IEEE.*

*[35]KRISHNAPPA, K. H., & Trivedi, S. K. (2023). Efficient and Accurate Estimation of Pharmacokinetic Maps from DCE-MRI using Extended Tofts Model in Frequency Domain.*

*[36]Krishnappa, K. H., Shashidhar, R., Shashank, M. P., & Roopa, M. (2023, November). Detecting Parkinson's disease with prediction: A novel SVM approach. In 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE) (pp. 1-7). IEEE.*

*[37]Shashidhar, R., Balivada, D., Shalini, D. N., Krishnappa, K. H., & Roopa, M. (2023, November). Music Emotion Recognition using Convolutional Neural Networks for Regional Languages. In 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE) (pp. 1-7). IEEE.*

*[38]Madhura, R., Krishnappa, K. H., Manasa, R., & Yashaswini, K. P. (2023, August). Slack Time Analysis for APB Timer Using Genus Synthesis Tool. In International Conference on ICT for Sustainable Development (pp. 207-217). Singapore: Springer Nature Singapore.*

*[39]Krishnappa, K. H., & Gowda, N. V. N. (2023, August). Dictionary-Based PLS Approach to Pharmacokinetic Mapping in DCE-MRI Using Tofts Model. In International Conference on ICT for Sustainable Development (pp. 219-226). Singapore: Springer Nature Singapore.*

*[40]Krishnappa, K. H., & Gowda, N. V. N. (2023, August). Dictionary-Based PLS Approach to Pharmacokinetic Mapping in DCE-MRI Using Tofts Model. In International Conference on ICT for Sustainable Development (pp. 219-226). Singapore: Springer Nature Singapore.*

*[41]Madhura, R., Krutthika Hirebasur Krishnappa. et al., (2023). Slack time analysis for APB timer using Genus synthesis tool. 8th Edition ICT4SD International ICT Summit & Awards, Vol.3, 207–217. https://doi.org/10.1007/978-981-99-4932-8_20*

*[42]Shashidhar, R., Aditya, V., Srihari, S., Subhash, M. H., & Krishnappa, K. H. (2023). Empowering investors: Insights from sentiment analysis, FFT, and regression in Indian stock markets. 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE), 01–06. https://doi.org/10.1109/AIKIIE60097.2023.10390502*

*[43]Jayakeshav Reddy Bhumireddy, Rajiv Chalasani, Mukund Sai Vikram Tyagadurgam, Venkataswamy Naidu Gangineni, Sriram Pabbineedi, Mitra Penmetsa. Predictive models for early detection of chronic diseases in elderly populations: A machine learning perspective. Int J Comput Artif Intell 2023;4(1):71-79. DOI: 10.33545/27076571.2023.v4.i1a.169*

[44]HK, K. (2020). Design of Efficient FSM Based 3D Network on Chip Architecture. *INTERNATIONAL JOURNAL OF ENGINEERING*, *68*(10), 67-73.

[45]Krutthika, H. K. (2019, October). Modeling of Data Delivery Modes of Next Generation SOC-NOC Router. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1-6). IEEE.

[46]Ajay, S., Satya Sai Krishna Mohan G, Rao, S. S., Shaunak, S. B., Krutthika, H. K., Ananda, Y. R., & Jose, J. (2018). Source Hotspot Management in a Mesh Network on Chip. In *VDAT* (pp. 619-630).

[47]Nair, T. R., & Krutthika, H. K. (2010). An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System. *arXiv preprint arXiv:1001.3781*.

[48]Gopalakrishnan Nair, T. R., & Krutthika, H. K. (2010). An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System. *arXiv e-prints*, arXiv-1001.

[49]Krutthika H. K. & A.R. Aswatha. (2021). Implementation and analysis of congestion prevention and fault tolerance in network on chip. *Journal of Tianjin University Science and Technology, 54*(11), 213–231. https://doi.org/10.5281/zenodo.5746712

[50]Kuraku, Dr Sivaraju, et al. "Exploring how user behavior shapes cybersecurity awareness in the face of phishing attacks." *International Journal of Computer Trends and Technology* (2023).

[51]Kuraku, D. S., & Kalla, D. (2023). Impact of phishing on users with different online browsing hours and spending habits. *International Journal of Advanced Research in Computer and Communication Engineering*, *12*(10).

[52]Kalla, D., & Samaah, F. (2023). Exploring Artificial Intelligence And Data-Driven Techniques For Anomaly Detection In Cloud Security. *Available at SSRN 5045491*.

[53]Chandrasekaran, A., & Kalla, D. (2023). Heart disease prediction using chi-square test and linear regression. *Comput. Sci. Inform. Technol.*, *13*, 135-146.

[54]Kalla, D. (2022). AI-Powered Driver Behavior Analysis and Accident Prevention Systems for Advanced Driver Assistance. *International Journal of Scientific Research and Modern Technology (IJSRMT) Volume*, *1*.

[55]Rajiv, C., Mukund Sai, V. T., Venkataswamy Naidu, G., Sriram, P., & Mitra, P. (2022). Leveraging Big Datasets for Machine Learning-Based Anomaly Detection in Cybersecurity Network Traffic. *J Contemp Edu Theo Artific Intel: JCETAI/102*.

[56]Sandeep Kumar, C., Srikanth Reddy, V., Ram Mohan, P., Bhavana, K., & Ajay Babu, K. (2022). Efficient Machine Learning Approaches for Intrusion Identification of DDoS Attacks in Cloud Networks. *J Contemp Edu Theo Artific Intel: JCETAI/101*.

[57]Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2020). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

[58]Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340

[59]Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2022). Designing an Intelligent Cybersecurity Intrusion Identify Framework Using Advanced Machine Learning Models in Cloud Computing. *Universal Library of Engineering Technology*, (Issue).

[60]Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2022). Leveraging Artificial Intelligence Algorithms for Risk Prediction in Life Insurance Service Industry. *Available at SSRN 5459694*.

[61]Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Data Security in Cloud Computing: Encryption, Zero Trust, and Homomorphic Encryption. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(3), 70-80.

[62]Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Tyagadurgam, M. S. V. Efficient Framework for Forecasting Auto Insurance Claims Utilizing Machine Learning Based Data-Driven Methodologies. *International Research Journal of Economics and Management Studies IRJEMS*, *1*(2).

[63]Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., & Patchipulusu, H. H. S. (2022). Blockchain Technology in Supply Chain and Logistics: A Comprehensive Review of Applications, Challenges, and Innovations. *International Journal of Emerging Research in Engineering and Technology*, *3*(3), 99-107.

[64]Narra, B., Vattikonda, N., Gupta, A. K., Buddula, D. V. K. R., Patchipulusu, H. H. S., & Polu, A. R. (2022). Revolutionizing Marketing Analytics: A Data-Driven Machine Learning Framework for Churn Prediction. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(2), 112-121.

[65]Polu, A. R., Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., & Gupta, A. K. BLOCKCHAIN TECHNOLOGY AS A TOOL FOR CYBERSECURITY: STRENGTHS, WEAKNESSES, AND POTENTIAL APPLICATIONS.

[66]Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2022). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

[67]Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340