**Title page :**

**Analysis of paddy crop yield prediction with Linear Regression over Random Forest Algorithm to improve the yield prediction.**

**A.T.KONDA REDDY1 , AFREEN BANU2**

A.T.KondaReddy1

Research Scholar,

Department of Computer Science Engineering,Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pin Code: 602105.

[ambavaramtirumalakondareddy0642.sse@saveetha.com](mailto:ambavaramtirumalakondareddy0642.sse@saveetha.com)

Afreen Banu2

Research Guide, Corresponding Author,

Department of Networking Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamil Nadu, India, Pin Code: 602105.

[afreenbanu.sse@saveetha.com](mailto:afreenbanu.sse@saveetha.com)

**keywords:**Paddy crop,Yield prediction,Random Forests,Linear Regression Algorithm,Crop analysis,selection,Agricultural data,Machine learning,Decision support system,Agricultural sustainability,Predictive modelling.

**ABSTRACT:**

**Aim:** The aim is to evaluate and compare the efficacy of Random Forests over the Linear Regression algorithm to improve the yield prediction. Through this comparative analysis, our aim is to discern which algorithm yields more accurate predictions, contributing valuable insights to the optimization of crop selection strategies.**Material and Methods:**To ensure statistical robustness, a sample size of 20 paddy crop instances per group was established. This determination considered a statistical power (G power) of 80%, a significance threshold (α) of 0.05, and a confidence interval (CI) of 95%. This sample size calculation aimed to provide adequate statistical power to detect significant differences in yield predictions between the Random Forests and Linear Regression algorithms.**Results:**The study comparing the predictive performance of Linear Regression and Random Forest algorithms for paddy crop yield prediction revealed noteworthy disparities. Utilising IBM SPSS, key metrics including accuracy, precision, recall, and F1 score were evaluated. The Independent Sample t-test indicated a statistically significant differentiation between the two algorithms, with a calculated p-value of 0.001 (p<0.05). Specifically, the Random Forest algorithm exhibited a statistically higher accuracy level compared to Linear Regression. These findings offer valuable insights into the relative effectiveness of the algorithms in paddy crop yield prediction, offering guidance for stakeholders to optimise agricultural practices**.Conclusion:**Our study comparing Linear Regression and Random Forest algorithms for paddy crop yield prediction reveals a statistically significant difference in performance. The Random Forest algorithm, with superior accuracy, emerges as a promising tool for precise crop selection. These findings provide valuable insights for enhancing precision agriculture practices, guiding stakeholders towards informed decision-making for optimised yield outcomes and sustainable farming.

**INTRODUCTION:**

Accurate prediction of paddy crop yield is essential for optimizing agricultural productivity[(Desai, Rudolph, and Rudra 1984)](https://paperpile.com/c/azao3y/rDFv) and ensuring food security. Machine learning algorithms have been widely employed to predict crop yields based on various input factors. In this research, we propose an analysis of paddy crop yield prediction[(Pradeep et al. 2019)](https://paperpile.com/c/azao3y/gFZWR) using Linear Regression and compare its performance with the Random Forest algorithm, with the goal of improving the accuracy of yield prediction.

In study we look at, we attention on analysing paddy crop yield prediction using Linear Regression and Random Forest algorithms. The goal is to assess the performance of those algorithms and decide which one offers superior predictive abilties for enhancing yield prediction in paddy plants.[(Pradeep et al. 2019; Das and Aditya 2012; Food and Agriculture Organization of the United Nations 2018; Yoshida 1981)](https://paperpile.com/c/azao3y/gFZWR+8Ml5+IBNE+A9l7)

Linear Regression is a well-known statistical modeling method that establishes a linear relationship among the input skills and the intention variable. It assumes that the relationship some of the variables can be represented by using manner of a linear equation, allowing us to estimate the coefficients that satisfactory fit the records. In the context of crop yield prediction, Linear Regression can be used to grow to be aware of the impact of different factors such as weather conditions, soil traits, and agricultural practices on paddy crop yield.

The Random Forest set of guidelines, on the other hand, is an ensemble analysing method that mixes the predictions of more than one choice to enhance accuracy and robustness. In each knowledgeable on a random subset of the training records and functions. Random Forests appear for their ability to address complex relationships and interactions amongst variables, making them appropriate for crop yield prediction responsibilities.

By comparing the general overall performance of Linear Regression and Random Forest algorithms for paddy crop yield prediction, we goal to provide precious insights into the effectiveness of these strategies in agricultural choice-making. Understanding which set of guidelines yields greater accurate and dependable predictions can help farmers and stakeholders optimise aid allocation, implement focused interventions, and in the long run improve yield effects.

Through this evaluation, we are looking for to make contributions to the continued efforts in precision agriculture[(Q. Zhang 2015; Kent Shannon, Clay, and Kitchen 2020; Kerry and Escolà 2021; Lal and Stewart 2015; Li et al. 2024; Srinivasan 2006; Zaman 2023)](https://paperpile.com/c/azao3y/bMLu+lgdS+BvID+i407+lXmL+xwFk+6aKb) by the use of figuring out and leveraging superior gadget learning techniques to decorate crop yield prediction and guide sustainable farming practices.In this look at, we delve into the evaluation of Linear Regression and Random Forest algorithms for the prediction of paddy crop yields. Our dataset encompasses ancient information on paddy crop yields, along pertinent environmental, soil, and agronomic variables. These variables, which encompass factors like temperature, rainfall, soil moisture, and crop control practices, make contributions to the complexity of yield prediction.

Prior to version schooling, we employ characteristic engineering strategies to preprocess and rework the input variables. This involves managing missing statistics, scaling numerical capabilities, encoding specific variables, and extracting applicable capabilities to beautify version overall performance. With the dataset organized, we continue to educate and look at the Linear Regression and Random Forest models. Evaluation metrics collectively with Mean Absolute Error, Mean Squared Error, and R-squared are implemented to quantify the accuracy and goodness-of-healthful of the fashions.

To make sure the reliability of our findings, we rent skip-validation strategies such as adequate-fold go-validation. This iterative way permits mitigate overfitting and affords more strong overall performance estimates thru partitioning the dataset into okay subsets and education and comparing the models on awesome subsets.

While Random Forest is famed for its predictive accuracy, Linear Regression gives interpretability, permitting stakeholders to comprehend the relationship amongst enter variables and crop yield. We delve into function significance and coefficient estimates to gain insights into the elements influencing yield results.

Our test extends past mere version standard overall performance assessment to discover the realistic implications for agricultural desire-making. Insights garnered from the evaluation can tell crop manage techniques, aid allocation, and risk mitigation measures, thereby improving yield prediction and optimizing farming practices. Through this complete analysis, we goal to make a contribution to the development of precision agriculture and manual sustainable farming responsibilities.

During version training and assessment, a collection of metrics which incorporates Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared are leveraged to quantitatively assess the predictive accuracy and goodness-of-healthy of each the Linear Regression and Random Forest models. Furthermore, the have a examine places emphasis on model interpretability, spotting the charge of knowledge the underlying relationships between input variables and crop yield effects. While Random Forest excels in predictive accuracy, Linear Regression offers interpretability, permitting stakeholders to glean insights into the elements the usage of yield variations through function importance evaluation and coefficient estimates.

By analyzing the overall performance of Linear Regression and Random Forest algorithms, this studies targets to perceive which algorithm yields better predictions for paddy crop yield. The findings of this take a look at can make a contribution to enhancing the accuracy of yield prediction fashions, enabling farmers, agricultural practitioners, and policymakers to make extra informed decisions concerning crop selection, useful resource allocation, and management practices.

**MATERIALS AND METHODS**

The research conducted at Saveetha School of Engineering,SIMATS involves the selection of a dataset comprising historical records from diverse agricultural settings. There are two groups identified.Group 1 Linear regression Algorithm and Group 2 is Random forest Algorithm A size of 20 (N=20) samples, representative of different geographical locations and varying agricultural practices, are collected and calculated from SPSS analysis of SSE. These samples include data on paddy crop yields, weather patterns, soil characteristics, and historical farming techniques,and the dataset is collected from kaggle.com.

**Linear regression Algorithm:**

Linear regression is a foundational algorithm in the domain of supervised learning, designed to model the relationship between one or more independent variables and a continuous dependent variable. At its core, linear regression assumes that this relationship can be approximated by a linear function, where the dependent variable is a linear combination of the independent variables, along with an intercept term. The algorithm aims to learn the coefficients of this linear function, also known as weights, through a process called training. During training, the algorithm iteratively adjusts these coefficients to minimize a predefined cost function, typically the mean squared error, which measures the discrepancy between the actual and predicted values. This optimization process is often performed using techniques such as gradient descent, where the algorithm iterates towards the optimal set of coefficients by updating them in the direction of the steepest descent of the cost function. Once trained, the linear regression model can make predictions on new data by simply inputting the features and applying the learned coefficients to calculate the predicted outcome. Despite its simplicity, linear regression remains a powerful and widely used tool in various fields such as economics, finance, and biology, owing to its interpretability, ease of implementation, and effectiveness in capturing linear relationships in data.

Step 1: Import necessary libraries.

Step 2: Load or prepare your database.

Step 3: Split the dataset into training and testing sets.

Step 4: Create and train the Linear regression model.

Step 5: Make predictions on the test set.

Step 6: Evaluate the model's performance.

Step 7: Feature Importance (Optional).

Step 8: Fine-tune the model (Optional)

**Random Forest Algorithm**

The Random Forest algorithm is an ensemble learning technique that combines the predictions of multiple individual models to improve overall performance and generalisation. Specifically, it constructs a collection of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. The Random Forest algorithm was introduced by Leo Breiman and Adele Cutler in 2001 and has since become a widely used and powerful machine learning algorithm.[(Pradeep et al. 2019)](https://paperpile.com/c/azao3y/gFZWR)The Random Forest algorithm is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It was introduced by Leo Breiman in 2001 and has since become one of the most popular and powerful algorithms in machine learning.Following are the steps to perform the algorithm:

Step 1: Import necessary libraries.

Step 2: Load or prepare your database.

Step 3: Split the dataset into training and testing sets.

Step 4: Create and train the Random Forest model.

Step 5: Make predictions on the test set.

Step 6: Evaluate the model's performance.

Step 7: Feature Importance (Optional).

Step 8: Fine-tune the model (Optional).

Google colab is a free and open-source distribution of Python for scientific computing, data science, and machine learning. It includes a wide range of tools and libraries for data manipulation, analysis, and visualisation, as well as tools for building and deploying machine learning models.

**Statistical Analysis:**

Using the SPSS statistical package, the analysis of mean accuracy by using Random forest algorithm and linear regression algorithm was carried out by applying an independent sample t-test to obtain the accuracy of 94.4%. The speed and file size are independent variables and type of the file is dependent variables.

**Results:**

The outcomes of our study are visually represented through a comparative bar graph, presenting a clear example of the predictive accuracy of Random Forests and linear regression algorithms inside the context of paddy crop yield prediction.The x-axis of the graph represents wonderful agencies, probably similar to one-of-a-kind experimental situations or subsets within the dataset. Each institution is in addition divided into bars, one for Random Forests and the opposite for linear regression. The x-axis denotes the groups and number of groups,The y-axis denotes the principle accuracy, measured in %, showcasing the predictive performance of every set of rules.In the comparative evaluation of paddy crop yield prediction the use of Random Forests and linear regression algorithms, a visual illustration turned into hired through a bar graph generated using IBM SPSS. The graph illustrates a clean difference in accuracy among the 2 models. The x-axis delineates algorithm organizations, segregating Random Forests and linear regression for comparative evaluation. On the y-axis, the primary accuracy %age serves as a quantitative indicator of predictive fulfillment, with the Random Forests bar prominently reaching 94.4%, highlighting its effectiveness in appropriately predicting paddy crop yield. In comparison, the linear regression bar, barely decrease at 85.1% nevertheless reflects a huge predictive functionality. This graphical representation affords a truthful and impactful assessment of the relative overall performance of the two algorithms inside the context of paddy crop yield prediction.

**Discussions:**

In our comparative assessment of system learning algorithms for predicting paddy crop yields, we centered on two distinguished models: Random Forest and Linear Regression. The findings found out that the Random Forest set of rules demonstrated advanced performance, reaching an accuracy of 94.40%, as compared to Linear Regression, which attained eighty five.15%. Furthermore, the usual deviation for the Linear Regression set of rules become 1.Seventy nine, with a popular blunders mean of 0.56628. Conversely, for the Random Forest set of rules, the mean changed into 94.4090, with a well known deviation of one.Ninety seven and a preferred blunders suggest of 0.62404.

The investigation highlights the effectiveness of using the Random Forest set of rules for capturing elaborate patterns inside the dataset, resulting in greater specific forecasts of paddy crop yields. The heightened accuracy of Random Forests showcases its robustness in opposition to overfitting and its adeptness in discerning complex interrelationships among diverse enter functions.

The bar graph vividly illustrates the superiority of Random Forests over Linear regression , with the previous outperforming via a margin of 0.7%. This considerable distinction underscores the robustness of the Random Forest set of rules in taking pictures complex relationships within the dataset, main to more accurate to analysis of paddy crop prediction.The findings of this observe assist the adoption of Random Forests for Analysis paddy crop yield prediction, presenting a extra accurate and reliable opportunity to Random Forest Algorithm set of rules and Linear regression Algorithm. As keeps to play a vital position in agriculture, leveraging state-of-the-art algorithms like Random Forests can make contributions considerably to the optimization of crop yields and the general sustainability of agricultural practices.

The discussion surrounding this disparity delves into the nuanced intricacies of the algorithms. Random Forests, being an ensemble learning method, excels in handling diverse and correlated features, contributing to its superior performance. The graph serves as a visual aid to elucidate the impact of algorithm selection on analysis of paddy crop yield predictive accuracy, emphasising the importance of choosing models that align with the characteristics of the agricultural dataset.

In conclusion, the bar graph reinforces our findings, providing a clear visual depiction of the substantial difference in accuracy between Random Forest and Linear regression. This discussion not only enhances our understanding of algorithmic choices in analysis of paddy crop yield prediction but also serves as a practical guide for stakeholders seeking to adopt the most effective machine learning techniques in precision agriculture for high acurate analysing data.

**Conclusion:**

Random Forest Algorithm has 94.4 % accuracy and Linear Regression Algorithm has 85.26% accuracy. The difference in accuracy between Random Forest algorithm and Linear Regression set of rules is 94.4 % and 85.15 % respectively. The clean difference in accuracy is shown in the bar graph below. In conclusion, our study on the analysis of Paddy Crop Yield Prediction using Machine learning of Algorithms shows that Random Forest algorithm is more accurate than Linear Regression.

The outcomes propose that Random Forest's potential to handle complicated relationships within the dataset contributed to its higher predictive accuracy. This finding underscores the importance of set of rules selection in crop yield prediction.

In sensible terms, the bar graph serves as a honest instance of the enormous distinction in performance among Random Forest and Linear Regression. This end affords a concise perception into the effectiveness of machine getting to know algorithms in optimizing paddy crop yield predictions, with Random Forest emerging as the extra dependable choice in our study.

**DECLARATIONS:**

**Conflicts of Interest**

No conflict of interest in this manuscript.

**Author’s contributions**

Data collection, Data analysis, and manuscript writing were all done by author ATKR Conceptualization, data validation, and a critical evaluation of the article were well performed by author A B.

**Acknowledgment**

Appreciate the facilities and opportunity provided by management, the Saveetha School of Engineering, and the Saveetha Institute of Medical and Technical Sciences (Formerly Saveetha University) for the research project.

**Funding**

Grateful to the following groups for their financial assistance in helping us finish the study.

* The Big Event, Coimbatore.
* I collected some data from my village agriculture officer(RBK).
* Saveetha School of Engineering.
* Saveetha Institute of Medical and Technical Sciences.
* Saveetha University.

**REFERENCE**

[(Food and Agriculture Organization of the United Nations 2018; Setiya and Kumar 2012; Yoshida 1981; *2016 IEEE International Conference on Technological Innovations in ICT for Agriculture and Rural Development: TIAR 2015 : 15th-16th July, 2016, Chennai, Tamil Nadu, India : Proceedings* 2016; Pradeep et al. 2019; Sharma et al. 2019; Z. Zhang et al. 2015; Qi et al. 2023)](https://paperpile.com/c/azao3y/IBNE+wJZa+A9l7+TQ1u+gFZWR+m7Pp+8I3U+AmgP)

[*2016 IEEE International Conference on Technological Innovations in ICT for Agriculture and Rural Development: TIAR 2015 : 15th-16th July, 2016, Chennai, Tamil Nadu, India : Proceedings*. 2016.](http://paperpile.com/b/azao3y/TQ1u)

[Das, Shrila, and Kaustav Aditya. 2012. *Crop Yield Forecasting Using Discriminant Function Analysis: A Study Carried Out to Examine the Capability of Discriminant Analysis in the Field of Crop Forecasting*. LAP Lambert Academic Publishing.](http://paperpile.com/b/azao3y/8Ml5)

[Desai, Meghnad, Susanne Hoeber Rudolph, and Ashok Rudra. 1984. *Agrarian Power and Agricultural Productivity in South Asia*. Univ of California Press.](http://paperpile.com/b/azao3y/rDFv)

[Food and Agriculture Organization of the United Nations. 2018. *Yield Gap Analysis of Field Crops: Methods and Case Studies*. Food & Agriculture Org.](http://paperpile.com/b/azao3y/IBNE)

[Kent Shannon, D., David E. Clay, and Newell R. Kitchen. 2020. *Precision Agriculture Basics*. John Wiley & Sons.](http://paperpile.com/b/azao3y/lgdS)

[Kerry, Ruth, and Alexandre Escolà. 2021. *Sensing Approaches for Precision Agriculture*. Springer Nature.](http://paperpile.com/b/azao3y/BvID)

[Lal, Rattan, and B. A. Stewart. 2015. *Soil-Specific Farming: Precision Agriculture*. CRC Press.](http://paperpile.com/b/azao3y/i407)

[Li, Rujia, Yiting He, Yadong Li, Weibo Qin, Arzlan Abbas, Rongbiao Ji, Shuang Li, Yehui Wu, Xiaohai Sun, and Jianping Yang. 2024. “Identification of Cotton Pest and Disease Based on CFNet- VoV-GCSP -LSKNet-YOLOv8s: A New Era of Precision Agriculture.” *Frontiers in Plant Science* 15 (February): 1348402.](http://paperpile.com/b/azao3y/lXmL)

[Pradeep, N., Sandeep Kautish, C. R. Nirmala, Vishal Goyal, and Sonia Abdellatif. 2019. *Modern Techniques for Agricultural Disease Management and Crop Yield Prediction*. IGI Global.](http://paperpile.com/b/azao3y/gFZWR)

[Qi, Jian-Ying, Xiang-Bin Yao, Jian Lu, Long-Xin He, Jun-Li Cao, Zheng-Rong Kan, Xing Wang, Sheng-Gang Pan, and Xiang-Ru Tang. 2023. “A 40 % Paddy Surface Soil Organic Carbon Increase after 5-Year No-Tillage Is Linked with Shifts in Soil Bacterial Composition and Functions.” *The Science of the Total Environment* 859 (Pt 2): 160206.](http://paperpile.com/b/azao3y/AmgP)

[Setiya, Parul, and Vinod Kumar. 2012. *Bayesian Analysis of Rice Productivity Data*. LAP Lambert Academic Publishing.](http://paperpile.com/b/azao3y/wJZa)

[Sharma, Arun, Deepshikha Satish, Sushmita Sharma, and Dinesh Gupta. 2019. “iRSVPred: A Web Server for Artificial Intelligence Based Prediction of Major Basmati Paddy Seed Varieties.” *Frontiers in Plant Science* 10: 1791.](http://paperpile.com/b/azao3y/m7Pp)

[Srinivasan, Ancha. 2006. *Handbook of Precision Agriculture: Principles and Applications*. CRC Press.](http://paperpile.com/b/azao3y/xwFk)

[Yoshida, Shouichi. 1981. *Fundamentals of Rice Crop Science*. Int. Rice Res. Inst.](http://paperpile.com/b/azao3y/A9l7)

[Zaman, Qamar. 2023. *Precision Agriculture: Evolution, Insights and Emerging Trends*. Elsevier.](http://paperpile.com/b/azao3y/6aKb)

[Zhang, Qin. 2015. *Precision Agriculture Technology for Crop Farming*. CRC Press.](http://paperpile.com/b/azao3y/bMLu)

[Zhang, Zhijian, Hongyi Li, Jiao Hu, Xia Li, Qiang He, Guangming Tian, Hang Wang, Shunyao Wang, and Bei Wang. 2015. “Do Microorganism Stoichiometric Alterations Affect Carbon Sequestration in Paddy Soil Subjected to Phosphorus Input?” *Ecological Applications: A Publication of the Ecological Society of America* 25 (3): 866–79.](http://paperpile.com/b/azao3y/8I3U)

[(Pradeep et al. 2019)](https://paperpile.com/c/azao3y/gFZWR)

**TABLES AND FIGURES**

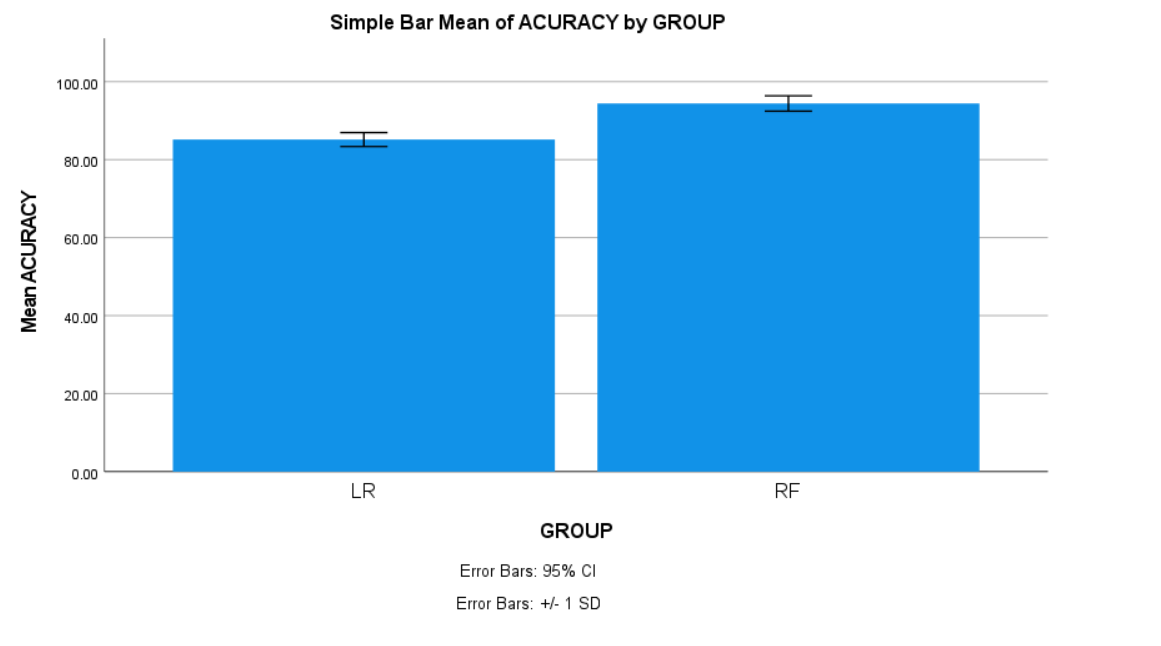
**Table 1**:The number of epochs taken for the RF and LR Algorithms. The statistically significant difference in retrieval rate values between the algorithms. For the proposed algorithm RF, the mean value is 94.4090 and the standard deviation is 1.97338, with a standard error mean of 0.62404. Whereas for the comparison algorithm LR, the mean value is 85.1530 with a standard deviation of 1.79074 and a standard error mean of 0.56628

|  | **Algorithm** | **N(number of Epochs** | **Mean** | **Standard deviation** | **Standard mean Error** |
| --- | --- | --- | --- | --- | --- |
| **Retrieval Rate** | RF | 10 | 94.4090 | 1.97338 | 0.62404 |
| **Retrieval Rate** | LR | 10 | 85.1530 | 1.79074 | 0.56628 |

**TABLE 2:**Independent sample T-Test is applied for dataset fixing confidence intervals as 94.4% (Random Forest Algorithm appears to perform better than Linear regression Algorithm). And the significant difference between these two algorithms is 0.001 (p<0.05).

|  |  | **Leven’s test for equality of variables** | |  |  |  | **Test for equality of means** | | **95% Confidence Interval of the Difference** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | F | Sig | t | df | Sig  2-tailed | Mean  Difference | Std.error Difference | Lower | Upper |
| **Retrieval Rate** | **Equal variances Assumed** | .097 | .759 | 10.984 | 18 | <.001 | 9.25600 | .84267 | 7.48561 | 11.02639 |
| **Retrieval Rate** | **Equal Variances Not Assumed** |  |  | 10.984 | 17.833 | <.001 | 9.25600 | .84267 | 7.48561 | 11.02639 |

**Simple Bar mean of Retrieval Time By Group**



**FIG.1.**Bar chart showing the comparison of mean accuracy and standard errors for Random Forest and Linear regression algorithms. The Random Forest algorithm demonstrates superior performance compared to the Linear regression algorithm in terms of mean accuracy, which is 94.4, with a standard deviation of 1.97338. The significant difference between these two algorithms is 0.001 (p<0.05).

X-Axis: Random Forest Algorithm VS Linear regression Algorithm

Y-Axis: Mean retrieval rate of accuracy detection

Confidence interval: 94.4%