2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) | 979-8-3315-2169-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/IATMSI64286.2025.10985667

Fake Profile Detection Using Machine Learning

S. N. Tirumala Rao
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
nagatirumalarao@gmail.com

Sireesha Moturi
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
sireeshamoturi@gmail.com

Suneetha Mothe
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
msuneetha973@gmail.com

Ravi Lakshmi Sri Harsha
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
raviharsha42@gmail.com

Najeer Shaik
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
nazeershaik4282@gmail.com

Sistla V. S. Manikanta Rohit
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
rohitsistla1212@gmail.com

Dodda Venkata Reddy
Dept of CSE,
Narasaraopeta Engineering College,
Narasaraopet-522601, Palnadu,
Andhra Pradesh, India
doddavenkatareddy@gmail.com

***Abstract*—In today's world, social media is an essential component of everyone's social life on the internet. It's become simpler to make new acquaintances and stay updated on their activities. Numerous fields are impacted by online social networks, including business, education, employment, community involvement, and research. Employers use these social networking sites to find and hire qualified applicants who are enthusiastic about their job. Spreading misinformation via social media is another problem. Incorrect accounts that propagate unsuitable and incorrect information may give rise to conflicts. These made-up accounts are likewise intended to attract followers. More harm is done to people by false profiles than by other internet crimes. Consequently, it's critical to be able to recognize a fake profile.**

***Index Terms*—Fake profile, Social media, Machine Learning, Gaussian Naïve Bayes classifier, Random Forest, Logistic Regression, XG(Extreme Gradient) Boosting, Fake profile detection**

I. Introduction

You have access to a wide range of contacts and possibilities via the Internet. Popular social networking services like Facebook, Instagram, Snapchat, and WhatsApp are certainly recognizable to you. Apart from these conventional modes of social engagement, the contemporary generation engages in a plethora of alternative types of interaction. Social networking services are quite simple to use for educators to instruct children and These days, teachers use these websites extensively to provide online lectures, give homework, host conversations, and more—all of which greatly enhance student learning. Employers may find candidates who are knowledgeable and passionate about their profession by using social networking sites. These websites make it easy to check candidates' backgrounds. While some of these platforms are free, others charge a membership fee, which they use for business purposes, and yet others rely on advertising to generate revenue. However, there are drawbacks, and false profiles are one of them. They often arise from a straightforward absence of in-person interactions, and this frequently results in invites that, in the absence of these phony [1] identities on social media, we wouldn't typically receive. Several studies in this field have been conducted due to the widespread usage of social networks. The majority of phony profiles are created to gain more followers by sending out spam and phishing attacks [2]. False accounts are outfitted with all the tools required for performing crimes online. False accounts pose a risk of data breaches and identity theft. False accounts purporting to be from individuals or groups may harm their reputation and acquire fewer likes and follows.

II. Literature Survey

The identification of malicious activities and fake accounts on social media has been a focal point of research, with numerous approaches proposed to tackle these issues.

Gururaj et al. [3] proposed using natural language processing (NLP) techniques to identify suspicious users based on their regular interactions. The authors highlighted anomalous behaviors to represent each user's activity patterns. They also demonstrated the effectiveness of support vector machine (SVM) classifiers for detecting harmful comments on blogs. This study emphasized the role of anomalous activities, behavior profiles, communications, and comment sections in identifying malicious users.

979-8-3315-2169-1/25/$31.00 ©2025 IEEE

Kodati et al. [4] introduced a novel hybrid classification method for detecting human-created fake profiles on social media platforms. The approach employs Spearman's rank-order correlation to reduce the feature vector, eliminating redundancy and selecting optimal features. The proposed hybrid SVM method achieved a 98% accuracy rate in identifying fake Twitter profiles, outperforming traditional machine learning approaches in terms of both performance and efficiency.

Hajdu et al. [5] applied artificial neural networks (ANN) and machine learning techniques to determine the likelihood of receiving a genuine Facebook friend request. Their work included an overview of relevant libraries, classes, and the sigmoid function, along with the calculation and application of weights. Additionally, the authors considered the specifications of social network pages, which are integral to their proposed solution.

Khaled et al. [6] proposed an innovative classification approach to improve the detection of fake accounts on social networks. Their methodology involved training an artificial neural network (ANN) using decision values obtained from an SVM model. The ”MIB” baseline dataset was preprocessed by applying several feature reduction techniques to optimize the feature vector. The hybrid ”SVM-NN” model demonstrated superior accuracy across all feature sets, achieving approximately 98% classification accuracy, significantly outperforming other classifiers.

These research efforts underscore the advancements in machine learning and hybrid techniques, enabling more effective detection of malicious behaviors and fake accounts on social platforms.

III. PROPOSED SYSTEM

Our Model is Proposed based on certain criteria as follows:

- Dataset Analysis
- Data Preprocessing Techniques [7]
- Creation and Evaluation of Model
- Acquisition of Model Accuracy

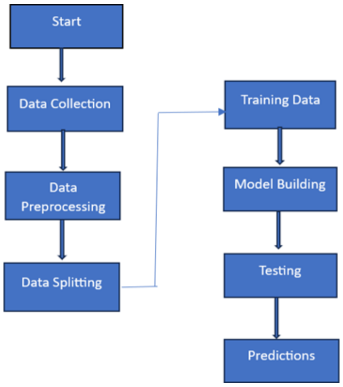
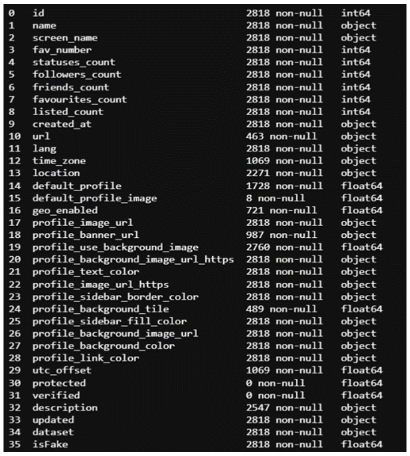


Fig. 1: Proposed Model

A. Data Analysis

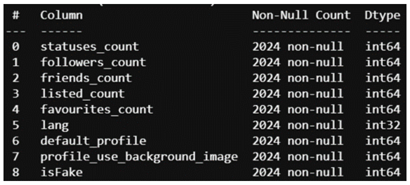
We have taken the datasets from the Kaggle website from the internet. We have collected two datasets which are users.csv and fusers.csv. The users and the fusers datasets contain 35 columns which is shown in Fig. 2. and it has both categorical and numerical data[8, 9].



```
0   id                                   2818 non-null   int64
1   name                                 2818 non-null   object
2   screen_name                          2818 non-null   object
3   fav_number                           2818 non-null   int64
4   statuses_count                       2818 non-null   int64
5   followers_count                      2818 non-null   int64
6   friends_count                        2818 non-null   int64
7   favourites_count                     2818 non-null   int64
8   listed_count                         2818 non-null   int64
9   created_at                           2818 non-null   object
10  url                                  463 non-null    object
11  lang                                 2818 non-null   object
12  time_zone                            1069 non-null   object
13  location                             2271 non-null   object
14  default_profile                      1728 non-null   float64
15  default_profile_image                8 non-null      float64
16  geo_enabled                          721 non-null    float64
17  profile_image_url                    2818 non-null   object
18  profile_banner_url                   987 non-null    object
19  profile_use_background_image         2760 non-null   float64
20  profile_background_image_url_https   2818 non-null   object
21  profile_text_color                   2818 non-null   object
22  profile_image_url_https              2818 non-null   object
23  profile_sidebar_border_color         2818 non-null   object
24  profile_background_tile              489 non-null    float64
25  profile_sidebar_fill_color           2818 non-null   object
26  profile_background_image_url         2818 non-null   object
27  profile_background_color             2818 non-null   object
28  profile_link_color                   2818 non-null   object
29  utc_offset                           1069 non-null   float64
30  protected                            0 non-null      float64
31  verified                             0 non-null      float64
32  description                          2547 non-null   object
33  updated                              2818 non-null   object
34  dataset                              2818 non-null   object
35  isFake                               2818 non-null   float64
```

Fig. 2: Dataset Description

These datasets tell us about the profile of each user, the user being either genuine or fake. Fig. 3. shows which data we are considering for further steps



```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   statuses_count                2024 non-null   int64
 1   followers_count               2024 non-null   int64
 2   friends_count                 2024 non-null   int64
 3   listed_count                  2024 non-null   int64
 4   favourites_count              2024 non-null   int64
 5   lang                          2024 non-null   int32
 6   default_profile               2024 non-null   int64
 7   profile_use_background_image  2024 non-null   int64
 8   isFake                        2024 non-null   int64
```

Fig. 3: Consider Data

B. Data PreProcessing Techniques

The first step in this study is data preprocessing, where raw data is transformed into a format suitable for analysis by the machine learning model. This involves various tasks such as data cleaning, normalization, and feature encoding. Raw data from social media platforms typically contains noise, missing values, and inconsistencies, which must be addressed before feeding it into the model. Additionally, preprocessing may involve extracting relevant features from the data that are indicative of fake profiles, such as activity patterns, profile completeness, and interaction behavior.

Machine learning algorithms do not allow missing values in the data, so handling missing values is crucial. Among the methods for dealing with missing values are: Swapping out for a mean, median, or mode, Back-fill or forward-fill, and Deleting row.

The next step in our study is to check whether the dataset has null values or not. The matplotlib library in Python provides a convenient way to visualize missing values [10] in a dataset. It offers various types of plots to help you understand the distribution and patterns of missing values within your dataset. Fig. 4. shows which columns have missing values.

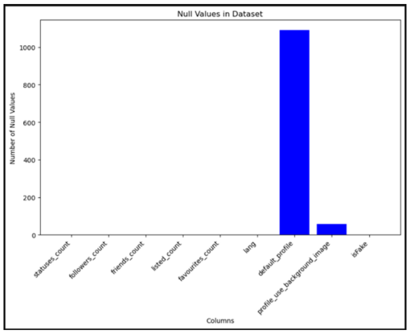


Fig. 4: Null values

Figure 5 depicts that there are no null values in our dataset and we have successfully removed all the null values from our datasets using the data preprocessing methods as mentioned above, we have removed the null values in our datasets.

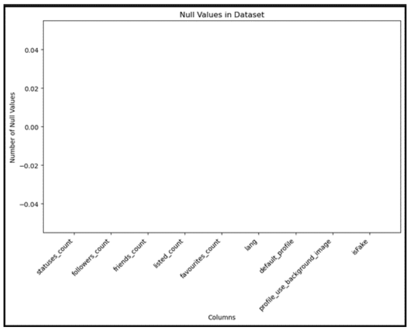


Fig. 5: After the removal of null values

Outlier detection is a crucial step in identifying anomalous instances within the dataset [11], which may signify the presence of fake profiles. Outliers could manifest as profiles exhibiting unusual activity levels, suspicious posting behavior, or inconsistencies in profile information. Leveraging outlier detection techniques, such as clustering-based approaches or statistical methods, enables the identification and subsequent handling of such instances. Outliers are detected and removed from the dataset as shown in Figures 6. and 7.

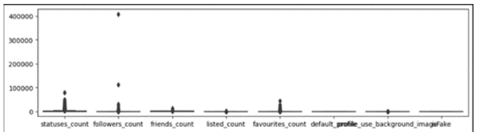


Fig. 6: Before Outlier Removal

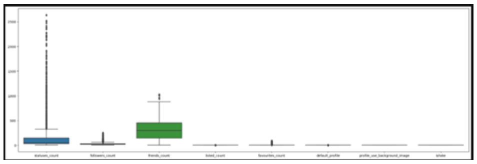


Fig. 7: After Outlier Removal

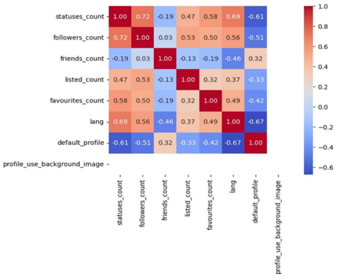


Fig. 8: Correlation Heatmap

A correlation heatmap [12, 13] visual representation of the dataset is plotted to get a brief understanding and visualization regarding which features are strongly correlated and which features are weakly correlated, as shown in Fig. 8.

In our dataset, the target variable is “isFake”. It contains the values of real users and fake users. Fig. 9. and Fig. 10. shows how the values are distributed. However, there is a class imbalance in Fig. 4. When there is value of a class is more than another value, it can lead to class imbalance. To remove class imbalance we are using SMOTE which is a technique used to address class imbalance in machine learning. The class is balanced after applying SMOTE [14], as shown in Fig. 5.

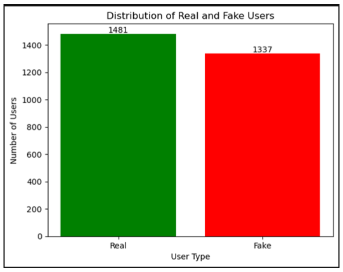


Fig. 9: Class imbalance

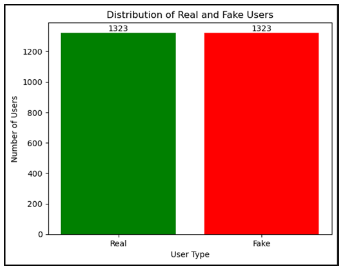


Fig. 10: Class Balance

C. Creation and Evaluation of Model

This step involves choosing an appropriate machine learning algorithm and training it on the ready data. To reduce the discrepancy between the model’s anticipated output and the actual output found in the training set, the parameters are optimized. The model is tested on a different validation dataset to gauge its performance after training. The performance criteria and problem type determine the evaluation measures that are employed. Evaluation criteria that are frequently used include F1 score, recall, accuracy, and precision. The model can be further improved by changing its parameters or using an alternative algorithm in light of the evaluation findings.

1) Feature Selection:: Feature selection is a crucial process [15] that enhances model performance and lowers computing costs by determining which subset of characteristics is most pertinent for categorization. It is crucial to use discriminative features for false profile identification that capture the unique traits that set real profiles apart from fraudulent ones. Finding the characteristics with the greatest predictive potential is made easier by methods including information gain, correlation analysis, and model-based selection.

2) Data Splitting:: The dataset is divided to assess how well the machine learning model performs. The model is trained using labeled data from the training set, which helps it discover the fundamental patterns that point to fraudulent profiles. The model's performance is then evaluated on the testing set, which consists of unobserved data. This assessment guarantees the model's good generalization to new cases and gives a precise evaluation of how effectively it detects phony profiles.

3) Model Evaluation: **Random Forest:**
The below Fig. 11 shows the pictorial representation of random forest.

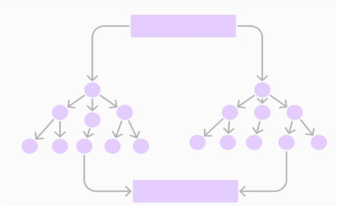


Fig. 11: Random Forest

Gradient Boosting:
Below Fig. 12 shows that combining different weaker learning models to construct a powerful prediction model.

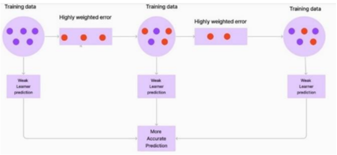
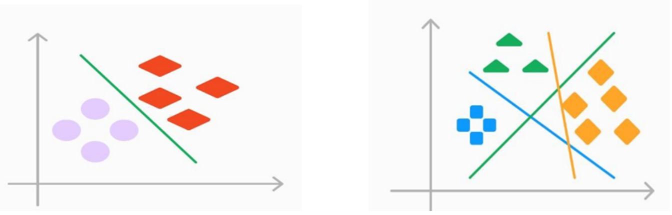


Fig. 12: Gradient Boosting

Logistic Regression:
Below Fig. 13 shows the classification between two classes, and Fig. 14 below shows the classification between multiple classes.



(a) Binary classification (b) Multiple classification

Fig. 13: Logistic Regression: Binary and Multiple Classifications

Gaussian Naive Bayes:

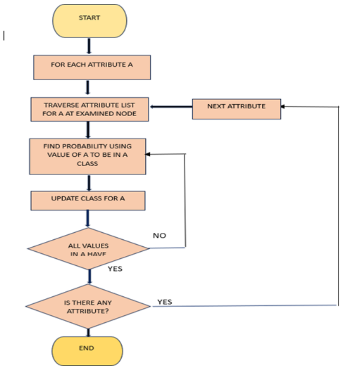


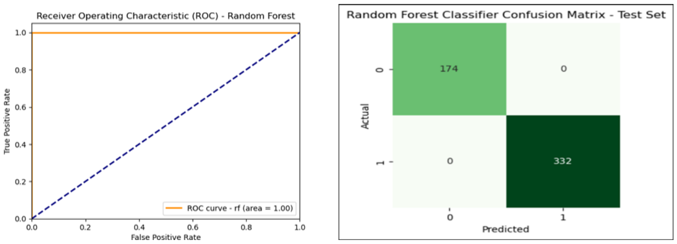
Fig. 14: Gaussian Naïve Bayes Algorithm

4) RESULT ANALYSIS: The models that we used in this study are Random Forest, Gradient Boosting Classifier, Logistic Regression, and Gaussian Naïve Bayes.

D. Random Forest Classifier

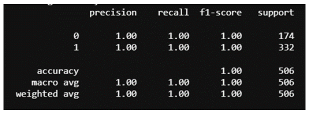
Random forest gives an accuracy of 100%. A ROC curve, which is shown below in Fig. 16. That the model can accurately categorize positive situations without producing any false positives, would look like a straight line in the graph's upper left corner.

The matrix[15] shows the performance of a random forest classifier on a test set. Fig. 17. indicates that the model made 174 true positive values, 332 true negative values, and 0 values for true negative and false positive. Fig. 18. and Fig. 19. Show the evaluation metrics of random forest.



(a) ROC Curve - Random Forest Classifier (b) Confusion Matrix of Random Forest

Fig. 15: Confusion Matrix and Metrics for Random Forest

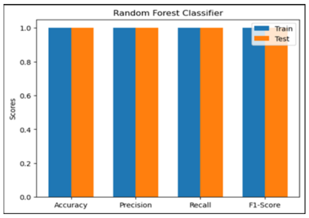


```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       174
           1       1.00      1.00      1.00       332

    accuracy                           1.00       506
   macro avg       1.00      1.00      1.00       506
weighted avg       1.00      1.00      1.00       506
```

(a) Metrics of Random Forest



(b) Metrics of Random Forest

Fig. 16: Metrics of Random Forest Comparisons

E. Logistic Regression

Logistic Regression gives an accuracy of 99%. The ROC Curve in Fig. 17 shows that the model is performing well.

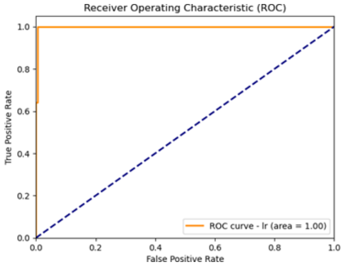
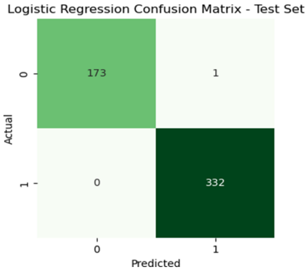
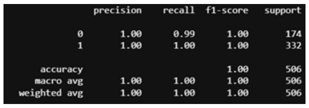


Fig. 17: ROC Curve Logistic Regression

The matrix shows the performance of Logistic Regression on a test set. Fig. 18a indicates that the model made 173 true positive values, 332 true negative values, 1 false negative, and 0 false positive.



(a) Confusion Matrix



```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       174
           1       1.00      1.00      1.00       332

    accuracy                           1.00       506
   macro avg       1.00      1.00      1.00       506
weighted avg       1.00      1.00      1.00       506
```

(b) Metrics for Logistic Regression

Fig. 18: Logistic Regression Evaluation

F. Naive Bayes

Naïve Bayes gives an accuracy of 99%. The ROC Curve in Fig. 19 shows that the model is performing well.

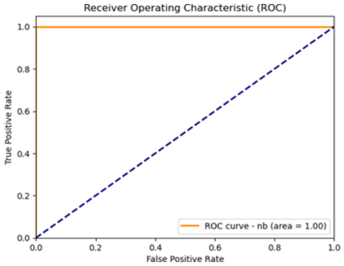
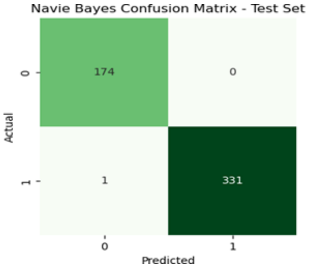
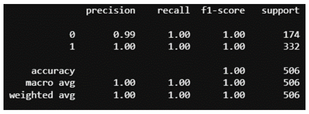


Fig. 19: ROC Curve Naïve Bayes

The matrix shows the performance of Naïve Bayes on a test set. Fig. 20a indicates that the model made 173 true positive values, 332 true negative values, 0 false negative, and 1 false positive.



(a) Confusion Matrix



```
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       174
           1       1.00      1.00      1.00       332

    accuracy                           1.00       506
   macro avg       1.00      1.00      1.00       506
weighted avg       1.00      1.00      1.00       506
```

(b) Evaluation Metrics 1

Fig. 20: Naive Bayes Evaluation

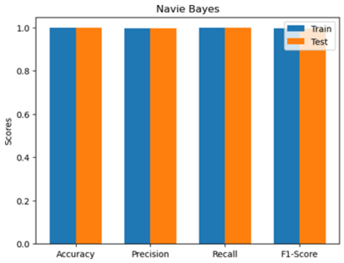


Fig. 21: Evaluation Metrics 2

G. Gradient Boosting Classifier

Gradient Boosting Classifier gives an accuracy of 99%. The ROC Curve in Fig. 22 shows that the model is performing well.

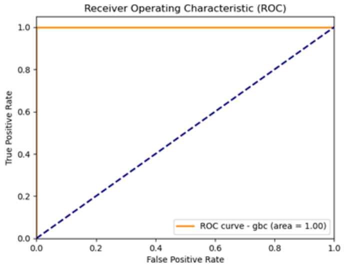
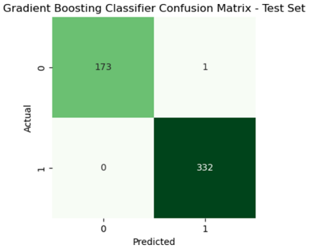
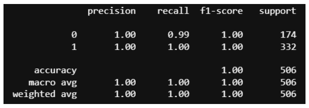


Fig. 22: ROC Curve Gradient Boosting

The matrix shows the performance of Gradient Boosting on a test set. Fig. 23a indicates that the model made 173 true positive values, 332 true negative values, 0 false negative, and 1 false positive.



(a) Confusion Matrix



```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       174
           1       1.00      1.00      1.00       332

    accuracy                           1.00       506
   macro avg       1.00      1.00      1.00       506
weighted avg       1.00      1.00      1.00       506
```

(b) Evaluation Metrics 1

Fig. 23: Gradient Boosting Evaluation

Once the model has been trained and evaluated, its efficiency and effectiveness in fake profile detection are analyzed. Performance metrics provide quantitative measures of the model's performance. we have used the random forest algorithm to achieve a higher accuracy of 100%, in the case of both the genuine and fake user accounts, as compared to the existing system which could achieve an accuracy score of 94% in the case of accurately detecting the genuine users and 97% in case of accurately classifying the fake users.

IV. Conclusion and Future Scope

This study explored the application of four popular machine learning algorithms, Random Forest, XG Boosting algorithm, and Logistic Regression, for predicting whether a social media account of a user is real or fake. The results indicate that of the four algorithms we have used for our predictions, the random forest algorithm is the most effective in generating accurate predictions, with random Forest outperforming the XG boost, Naïve Bayes, and Logistic Regression in terms of accuracy and efficiency. The predictive model developed in this study can be useful for better safety and privacy protection of authentic users from the fraudulent practices of fake users on social media platforms.

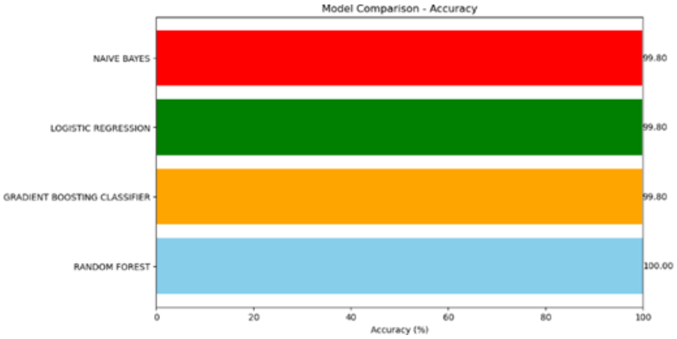


Fig. 24: Accuracy Comparision

Overall, this project highlights the potential of machine learning in enhancing user safety and privacy and allowing safe online engagement of users on social media platforms. Future research can expand this work by incorporating more complex features, exploring more optimized machine learning algorithms, and analyzing each machine learning model's performance on a larger dataset.

References

[1] Goyal, Bharti, Gill, Nasib, & Gulia, Preeti. (2023). Exploring machine learning techniques for fake profile detection in online social networks. *International Journal of Electrical and Computer Engineering*, 13, 2962-2971. https://doi.org/10.11591/ijece.v13i3.pp2962-2971.

[2] Harish, K., Kumar, R., & Bell J, Briso Becky. (2023). Fake Profile Detection Using Machine Learning. *International Journal of Scientific Research in Science, Engineering and Technology*, 719-725. https://doi.org/10.32628/IJSRSET2310264.

[3] Tanuja, U., Ramesh, B., H L, Gururaj., & Janhavi, V. (2021). Detecting malicious users in the social networks using machine learning approach. *International Journal of Social Computing and Cyber-Physical Systems*, 2, 229. https://doi.org/10.1504/IJSCCPS.2021.10041246.

[4] Kodati, Sarangam, Reddy, Kumbala, Mekala, Sreenivas, Murthy, P.L., & Sekhar reddy, Dr P Chandra. (2021). Detection of Fake Profiles on Twitter Using Hybrid SVM Algorithm. *E3S Web of Conferences*, 309, 01046. https://doi.org/10.1051/e3sconf/202130901046.

[5] Hajdu, Gergo, Minoso, Yaclaudes, Lopez, Rafael, Acosta, Miguel, & Elleithy, Abdelrahman. (2019). Use of Artificial Neural Networks to Identify Fake Profiles. In *2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, 1-4. https://doi.org/10.1109/LISAT.2019.8817330.

[6] Khaled, Sarah, El-Tazi, Neamat, & Mokhtar, Hoda. (2018). Detecting Fake Accounts on Social Media. *Proceedings of 2018 IEEE International Conference on Big Data*, 3672-3681. https://doi.org/10.1109/BigData.2018.8621913.

[7] M. Sireesha, S. N. Tirumala Rao, & Srikanth Vemuru. (2019). Optimized Feature Extraction and Hybrid Classification Model for Heart Disease and Breast Cancer Prediction. *International Journal of Recent Technology and Engineering*, 7(6), 1754-1772. ISSN: 2277-3878.

[8] Sireesha Moturi, Srikanth Vemuru, & S. N. Tirumala Rao. (2022). Two Phase Parallel Framework For Weighted Coalesce Rule Mining: A Fast Heart Disease And Breast Cancer Prediction Paradigm. *Biomedical Engineering: Applications, Basis and Communications*, 34(03). https://doi.org/10.4015/S1016237222500107.

[9] Joshi, U. D., Singh, A. P., Pahuja, T. R., Naval, S., & Singal, G. (2021). Fake Social Media Profile Detection. In: Srinivas, M., Sucharitha, G., Matta, A., & Chatterjee, P. (Eds.), *Machine Learning Algorithms and Applications*, Scrivener Publishing LLC, 193-209. https://doi.org/10.1002/9781119769262.ch11.

[10] Meshram, Pranay, Bhambulkar, Rutika, Pokale, Puja, Kharbikar, Komal, & Awachat, Anushree. (2021). Automatic Detection of Fake Profile Using Machine Learning on Instagram. *International Journal of Scientific Research in Science and Technology*, 117-127. https://doi.org/10.32628/IJSRST218330.

[11] Sunayna, S. S., Rao, S. N. T., & Sireesha, M. (2022). Performance Evaluation of Machine Learning Algorithms to Predict Breast Cancer. In: Nayak, J., Behera, H., Naik, B., Vimal, S., & Pelusi, D. (Eds.), *Computational Intelligence in Data Mining*, 281, 25. Springer, Singapore. https://doi.org/10.1007/978-981-16-9447-9$_{2}$5.

[12] S. L. Jagannadham, K. L. Nadh, & M. Sireesha. (2021). Brain Tumour Detection Using CNN. In: *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 734-739. https://doi.org/10.1109/I-SMAC52330.2021.9640875.

[13] Moturi, S., et al. (2024). Prediction of Liver Disease Using Machine Learning Algorithms. In: Nanda, S. J., Yadav, R. P., Gandomi, A. H., & Saraswat, M. (Eds.), *Data Science and Applications. ICDSA 2023. Lecture Notes in Networks and Systems*, vol 820. Springer, Singapore. https://doi.org/10.1007/978-981-99-7817-5$_{1}$9.

[14] Mamidala, Sai, Sireesha, M., Rao, S., Bolla, Jhansi, & Reddy, K. (2024). Machine Learning Models for Chronic Renal Disease Prediction. In: *Proceedings of Data Science and Applications ICDSA 2023*, 173-182. https://doi.org/10.1007/978-981-99-7820-5$_{1}$4.

[15] Reddy, S. D. P. (2019). Fake Profile Identification Using Machine Learning. *International Research Journal of Engineering and Technology (IRJET)*, 6, 1145-1150.