IJITCE

ISSN 2347- 3657

# International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 11, Issue 2, 2023

# Detecting Fake Reviews Through the Integration of Machine Learning and NLP Techniques

*Mrs.Sujatha Godavarthi[1]., P.Sindhuja[2]*

*1 Assistant Professor, Department of CSE, Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India*
*2, B.Tech CSE (20RG1A0545),*
*Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India*

## ABSTRACT

The proliferation of online shopping platforms has brought about a surge in user-generated product reviews, making it susceptible to the infiltration of fake reviews. For these platforms to continue to be dependable and reliable, it is necessary to identify and mitigate the impact of fake reviews. When depending on reviews for the product present on various web pages and applications, the rate of false reviews has been growing in the e-commerce sector. The goal is to anticipate and identify fraudulent reviews on e-commerce sites, namely Amazon, by using a hybrid model that combines classic machine learning (ML) with natural language processing (NLP). The proposed hybrid approach that has been suggested seeks to improve detection accuracy and interpretability by utilizing the combined abilities of ML and NLP technologies. Our method combines the power of Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art language model and Bag of Words (BoW) with traditional ML algorithms like Random Forest and XGBoost. To enhance model performance, we employ stacking ensemble method with logistic regression as the meta learner. The machine learns complex linguistic patterns, contextual information, and cooperative behaviors suggestive of fake reviews through training on many datasets. The outcomes of the experimental evaluations demonstrate the effectiveness of the hybrid model, surpassing existing methods in accuracy and robustness. This research contributes to a reliable solution, poised to enhance the trustworthiness of online product reviews and fortify consumer decision-making processes which guarantees continued safety and assurance in online shopping environments

## 1. INTRODUCTION

### 1.1Background

The rise of the digital era has fundamentally transformed how consumers make purchasing decisions. Central to this transformation is the proliferation of online reviews, which provide valuable insights into the quality and performance of products and services. Online reviews have become an essential part of e-commerce, influencing consumer behavior and decisions. As a result, they have become a focal point for both businesses and customers in evaluating the credibility and popularity of products. However, the integrity of online reviews is increasingly threatened by the proliferation of fake reviews. Fake reviews are deceitful entries that aim to manipulate the perceptions of potential customers. They can be positive reviews meant to unfairly boost a product’s reputation or negative reviews intended to damage a competitor’s image. These reviews undermine the trustworthiness of online feedback mechanisms and mislead consumers, leading to potential financial losses and diminished consumer trust in online platforms. Identifying fake reviews is a significant challenge due to their often sophisticated nature. Unlike spam or blatant fraud, fake reviews are crafted to appear genuine and convincing. They can mimic the tone, style, and content of legitimate reviews, making them difficult to detect through simple rule-based systems. The dynamic and evolving strategies employed by those who generate fake reviews further complicate detection efforts. Thus, there is a growing need for advanced techniques that can accurately identify and filter out fake reviews, preserving the credibility of online review systems.

### 1 1.2 Motivation

The proliferation of online reviews has fundamentally transformed how consumers make purchasing decisions. Reviews provide valuable insights into product quality, service reliability, and overall user satisfaction, significantly influencing consumer behavior. However, this system's integrity is increasingly threatened by the growing problem of fake reviews—fabricated evaluations designed to deceive potential buyers for various motives, including financial gain, competitive sabotage, or promotional activities. Fake reviews can severely undermine trust in online platforms, leading to misinformation. dissatisfaction, and distorted market competition.

### 1.2 Importance of Fake Review Detection

In the digital age, online reviews are pivotal in consumer decision-making. Platforms like Amazon, Yelp, and TripAdvisor host millions of reviews that influence purchasing choices. However, fake reviews—whether positive or negative—undermine the credibility of these platforms, leading to consumer deception and financial loss. Nearly 70% of people rely on the internet for daily necessities, and they often depend solely on product ratings and reviews for their decisions. Fake reviews aim to mislead consumers, causing them to buy or avoid certain products without an effective system to distinguish genuine from false evaluations. Understanding the nuances between original and computer-generated reviews is crucial for developing effective detection mechanisms. Original reviews are typically written by actual customers who have interacted with the product or service. They often contain personal anecdotes, specific details about the user experience, and unique insights. The language used in genuine reviews may vary in tone, style, and vocabulary, reflecting the diversity of genuine customer feedback. On the other hand, computer-generated reviews are created using algorithms or templates, often by entities with an interest in manipulating product ratings. These reviews may lack specific details and a personal touch, often appearing generic or overly polished. They tend to follow repetitive patterns, with similar phrases and sentence structures being used frequently. 1.3 Objective The primary objective of this project is to develop a comprehensive system for the accurate identification of fake reviews through a hybrid fusion of machine learning and natural language processing techniques. This involves creating a robust model that can effectively discern genuine reviews from deceptive ones by leveraging the strengths of both ML and NLP.

## 2. LITERATURE SURVEY

### 2.1 Introduction to Fake Reviews

Fake reviews, also known as deceptive reviews or opinion spam, are deliberately false or misleading assessments of products or services. Their primary purpose is to artificially enhance or damage a product's reputation, influencing consumer decisions and market dynamics. Fake reviews can be classified into several types, including positive fake reviews intended to promote products, negative fake reviews designed to tarnish competitors, and incentivized reviews, where reviewers are paid or rewarded for their input. The prevalence of fake reviews has significant ramifications for both businesses and consumers. For businesses, fake reviews can distort perceived customer satisfaction, leading to unfair competitive advantages or disadvantages. For consumers, fake reviews undermine trust in online reviews, making it challenging to make informed purchasing decisions. As e-commerce continues to grow, the detection and mitigation of fake reviews have become critical to maintaining the integrity of online marketplaces. Detecting fake reviews is challenging due to their sophisticated nature and the constant evolution of deceptive strategies. Fraudsters employ various tactics, including complex language patterns, fake accounts, and coordinated review posting, to evade detection. This necessitates advanced and adaptive detection methods that can keep pace with evolving deception techniques.

### 2.2 Types of Fake Reviews

Fake reviews can be broadly categorized into the following types:

1. Positive Fake Reviews: These reviews aim to boost the reputation of a product or service. They often include exaggerated praise and high ratings to artificially inflate the product's perceived quality.

2. Negative Fake Reviews: These are designed to harm a competitor's product or service by providing negative feedback, low ratings, and highlighting non-existent flaws.

3. Incentivized Reviews: Reviewers are paid or given incentives to write reviews, which may not reflect their genuine opinions.

4. Unverified Purchase Reviews: Reviews from users who have not purchased the product, thus potentially lacking authenticity.

5. Bulk Reviews: Large volumes of reviews posted in a short time frame, often part of a coordinated effort to sway public opinion.

### 2.3 Existing Methods for Fake Review Detection

Current methods for detecting fake reviews employ a variety of techniques. Machine learning algorithms, such as Support Vector Machines (SVM) and Random Forests, analyze textual features and reviewer behaviors. Natural Language Processing (NLP) techniques, including sentiment analysis and topic modeling, help identify linguistic patterns indicative of deception. Deep learning models, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have shown promise in capturing complex data patterns.

#### 2.3.1 Machine Learning Approaches

Machine learning (ML) methods have been extensively applied to fake review detection, leveraging supervised, unsupervised, and semi-supervised techniques. In supervised learning, models are trained on labeled datasets to classify reviews as fake or genuine. Popular algorithms include Support Vector Machines (SVM) , Naive Bayes , Decision Trees , and ensemble methods like Random Forests . These models use a variety of features derived from review text. such as lexical, syntactic, and behavioral patterns, to identify deceptive reviews. Supervised approaches generally offer high accuracy but depend heavily on the quality and quantity of labeled data, which can be expensive and time-consuming to obtain. Unsupervised learning techniques do not require labeled data and aim to detect outliers or anomalies in the dataset. Clustering algorithms like K-means and anomaly detection methods such as Isolation Forests are commonly used. These approaches are beneficial for identifying novel types of review spam and can adapt to changing fraud tactics. However, they often suffer from lower accuracy compared to supervised models and may require significant tuning and validation to produce reliable results. Semi-supervised learning combines elements of both supervised and unsupervised methods, utilizing a small amount of labeled data alongside a larger pool of unlabeled data. This approach can balance the need for labeled data with the ability to detect new types of spam. While each of these methods has strengths and weaknesses, the ongoing challenge lies in improving their adaptability, scalability, and accuracy in real-world applications.

### 2.3.2 Natural Language Processing Techniques

Natural Language Processing (NLP) techniques are crucial in analyzing the textual content of reviews to identify fake ones. These methods encompass various stages of text analysis, from basic text processing to advanced semantic analysis. Sentiment analysis, for example, evaluates the sentiment expressed in reviews to detect discrepancies between the overall sentiment and specific content aspects. This can be particularly useful in spotting exaggeratedly positive or negative reviews that deviate from typical patterns. Linguistic features play a significant role in detecting fake reviews. Lexical features, such as word frequency and usage patterns, and syntactic features, like part-of-speech tags and sentence structure, can indicate unnatural language use typical in fake reviews . Semantic analysis goes deeper, focusing on the meaning of words and their context within the text. Techniques like Latent Dirichlet Allocation (LDA) for topic modeling and the use of word embeddings (e.g., Word2Vec, GloVe, BERT) enable the capture of semantic relationships and underlying themes in reviews, which helps in identifying deceptive content. Behavioral features, including review patterns and user behavior, further enhance detection capabilities. Analysis of metadata, such as the timing and frequency of reviews, and the reviewer's history, provides additional clues about potential fakery multifaceted approach to detecting fake reviews, balancing linguistic, semantic, and behavioral insights to improve accuracy and robustness.

### 2.4 Summary of Related Work

There are typically a lot of different ways to identify phoney evaluations, but it's challenging to match modern technologies while maintaining accuracy. This paper aims to enhance sentiment analysis for Indonesian product reviews by combining machine learning and deep learning models. They utilize TF-IDF, Word2Vec, logistic regression, SVC, and IndoBERT's pretrained model, assessing performance with accuracy, precision, recall, and F1-Score. While machine learning leads in precision, deep learning excels in accuracy, recall, and F1-Score . This research focuses on identifying fake reviews using sentiment analysis and natural language processing. It employs deep learning neural networks such as GRU, Bi-LSTM, and LSTM, alongside activation functions like ReLu, TanH, and Sigmoid, to analyze review feedback. Pre-processing techniques are used to transform data for effective analysis and detection . This paper addresses the issue of identifying and filtering fake reviews to improve the reliability of online reviews. It focuses on designing a machine learning model for fake review detection and compares the performance of three algorithms. This research focuses on employing supervised machine learning techniques to detect 15 and filter fake reviews, aiming to uphold the credibility of online reviewing systems. Future improvements could involve refining the preprocessing techniques, exploring ensemble learning methods. Focuses on deep learning methodologies, CNN and LSTM for the detection of fake reviews in online platforms. The study demonstrates the superior accuracy of deep learning models. However, to enhance this approach further, future research could explore ensemble techniques combining deep learning with traditional methods for more robust fake review detection. This research addresses the challenge of detecting fake online reviews by developing a comprehensive model based on ten psychological deception theories and nine relevant constructs. Using features extracted from Yelp datasets, the model was empirically validated with machine learning algorithms, demonstrating the importance of both verbal and non-verbal features. The theory-based model outperformed existing detection models, offering high interpretability and low complexity. This paper focuses on the SVM method to find fake reviews. It incorporates sentiment analysis to divide reviews into real and fake groups, filtering out false ones and recommending genuine products to users. Future improvements could involve enhancing the SVM algorithm's robustness and exploring additional features such as user behavior analysis to further improve the accuracy of fake review detection. T. -Y. Lin, B.

## 3. SYSTEM DESIGN

### 3.1 Overall Architecture

The system integrates various stages into a continuous pipeline, ensuring seamless data flow from collection through preprocessing, feature extraction, model training, and evaluation. Each stage is meticulously connected, facilitating a streamlined process where cleaned and normalized data progresses methodically to feature extraction, empowering models with robust inputs for training. This cohesive design not only enhances efficiency but also supports thorough evaluation, validating model performance across different subsets of the data and ensuring reliable classification of reviews as genuine or fake.

### 3.1.1 Data Flow Diagram

The data flow diagram (Figure 4.1) illustrates the flow of data through the system from data collection to prediction.

**1. Data Collection:** Reviews are sourced from publicly available datasets containing both genuine and computer-generated fake reviews. Each review includes features such as the review text, category, rating, and a label indicating its authenticity.

**2. Data Preprocessing:** The collected reviews undergo thorough cleaning and normalization processes. This involves removing HTML tags, converting text to lowercase, eliminating stop words, tokenizing the text into individual words, and lemmatizing words to reduce them to their base forms. These steps ensure that the text data is uniform and ready for feature extraction.

**3. Feature Extraction:** The cleaned reviews are processed to extract meaningful features using Bag-of-Words (BoW) and BERT embeddings. BoW transforms the text into a sparse matrix of token counts, capturing the frequency of each word in the corpus. BERT embeddings, on the other hand, provide context-aware semantic 23 representations of the review text, encoding nuanced meanings and relationships between words.

**4. Hybrid Feature Fusion:** To leverage both traditional statistical methods and advanced contextual understanding, the BoW and BERT embeddings are concatenated into a hybrid feature set. This fusion aims to enhance the models' ability to distinguish between genuine and fake reviews by incorporating both frequency-based and contextual information.

**5. Model Training:** The hybrid feature set is used to train models such as Random Forest and XGBoost. A stacking ensemble model is also trained, which uses Logistic Regression to combine the outputs of the base models. This ensemble approach improves predictive accuracy by integrating diverse model perspectives.

**6. Model Validation:** Models are evaluated using metrics like accuracy, precision, recall, F1- score, and AUC to ensure robustness and effectiveness in classifying fake reviews.

**7. Prediction:** When presented with a new review, the system follows a standardized pipeline. The review undergoes preprocessing steps like cleaning, stop word removal, tokenization, and lemmatization. Features are then extracted using both BoW and BERT embeddings, forming a hybrid feature vector that captures both statistical and contextual information. The trained models then classify the review as genuine or fake.

**8. Output:** Based on the hybrid feature representation, the trained models classify the review as either genuine or fake. Alongside the classification result, the system provides confidence scores or probabilities, indicating the certainty of the model's prediction. This information helps users interpret the reliability of the classification output

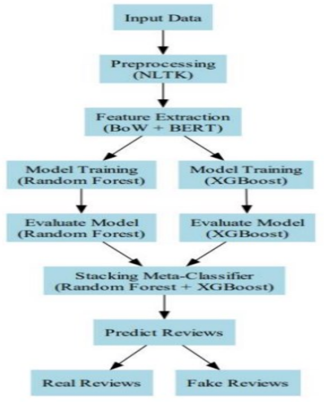

Input Data
Preprocessing (NLTK)
Feature Extraction (BoW + BERT)
Model Training (Random Forest)
Model Training (XGBoost)
Evaluate Model (Random Forest)
Evaluate Model (XGBoost)
Stacking Meta-Classifier (Random Forest + XGBoost)
Predict Reviews
Real Reviews
Fake Reviews

Fig 3.1 Data Flow Diagram

### 3.2 Model Training

The dataset for fake review detection is divided into training and testing sets, with 80% of the data allocated for training and 20% for testing. To prepare the data for model training, a hybrid feature set is created by combining Bag of Words (BoW) and BERT embeddings. This hybrid representation leverages the strengths of both traditional and modern text representation techniques, providing a comprehensive understanding of the review text. Each machine learning model Random Forest, XGBoost, and the stacking ensemble is trained on the training set. The Random Forest model builds multiple decision trees and aggregates their outputs for robust predictions. XGBoost, a scalable gradient boosting method, constructs decision trees sequentially, focusing on correcting errors from previous iterations. The stacking ensemble integrates predictions from both base models (Random Forest and XGBoost) using Logistic Regression as a meta-learner, aiming to enhance classification accuracy.

### 3.3 ACTIVITY DIAGRAM:

The Activity Diagram focuses on the dynamic aspects of the system, outlining the workflow for symptom input, processing, and recommendation generation. It visually represents the sequence of actions from the moment a user enters symptoms to when they receive personalized advice

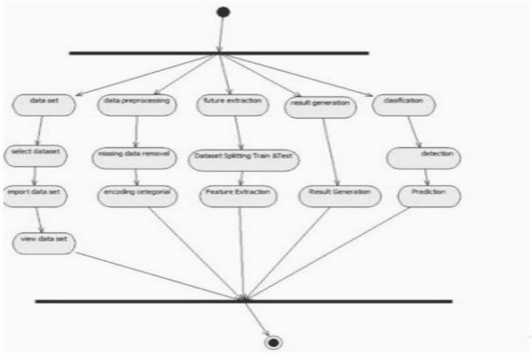


Fig 3.3 Activity Diagram

# 4. OUTPUT SCREENS

## 4.1 Experimental Setup

### 4.1.1 Dataset Details

The dataset used in this study consists of over 40,000 reviews, including both computer-generated fake reviews and genuine reviews provided by actual users. Both computer generated and real reviews are divided equally which is shown in fig. 5.1. The reviews were collected from various online sources and labeled to indicate whether they are real or fake.

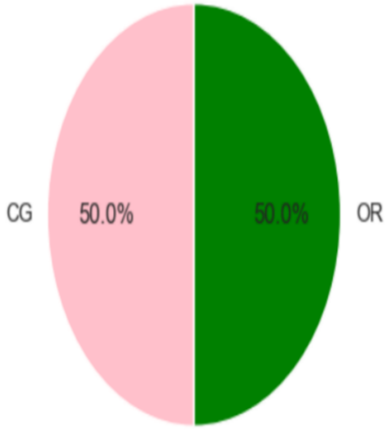


Figure 4.1: Computer generated and Original reviews count

FIG: 4.1 FINDING ACCURACY

For model training and evaluation, the dataset was split into training and testing sets. 80% of the data (approximately 32,000 reviews) was used for training, and 20% (around 8,000 reviews) was reserved for testing. This split ensures that the models are trained on a sufficient amount of data while being rigorously evaluated on unseen data to measure their generalizability.

## 4.2 Model Evaluation

### 4.2.1 Performance Metrics

To assess the performance of the models, we utilized several key metrics: Accuracy, which measures the proportion of correctly classified reviews; Precision, indicating the accuracy of positive predictions.

## 4.3 Comparative Analysis

### 4.3.1 Random Forest

The Random Forest classifier was tested using two feature extraction methods: Bag of Words (BoW) and BERT embeddings. The performance of Random Forest on these features was as follows:

- Random Forest (BoW): Achieved an accuracy of 80.66%. BoW captures the frequency of terms in the text, enabling the model to utilize word occurrence patterns to differentiate between fake and real reviews.

- Random Forest (BERT): Achieved an accuracy of 79.26%. BERT embeddings provide contextual understanding of the text, which helps in understanding the semantic content of the reviews.

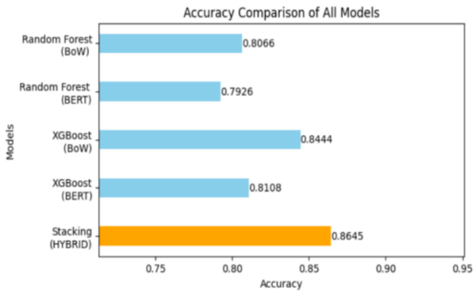
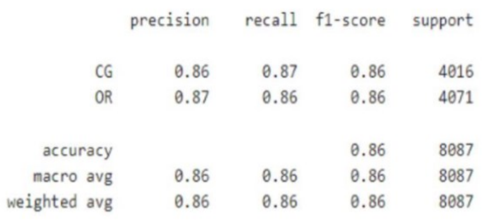


Figure 4.2: Accuracy Graph showing all models

### 4.4.1 Classification Report

The Stacking Classifier with BoW, BERT and traditional ML models achieved the precise accuracy of 86.45%, surpassing other techniques. A detailed classification report for the Stacking Classifier is shown in fig. 5.3. The weighted average metrics reflect the overall robust performance of this model, making it the most effective for fake review detection.



```
              precision    recall  f1-score   support

          CG       0.86      0.87      0.86      4016
          OR       0.87      0.86      0.86      4071

    accuracy                           0.86      8087
   macro avg       0.86      0.86      0.86      8087
weighted avg       0.86      0.86      0.86      8087
```

Figure4.3: Classification Report

## 5. CONCLUSION

In conclusion, our method illustrates the efficiency of a hybrid fusion combining NLP and ML models to enhance the accuracy and reliability of automated false review identification systems. Our research trained multiple machine learning (ML) classifiers, such as Random Forest and XGBoost and combined BoW and BERT embeddings. Our findings suggest that leveraging both word frequency and contextual embeddings can significantly increase the detection of fake reviews when combined using a stacking meta-classifier. Comparing our methodology against individual classifiers and baseline methods, the experimental results showed how successful it is in enhancing classification efficiency and dependability. The higher performance metrics were a result of the synergy between the predictive power of ensemble learning and BERT's semantic knowledge. By achieving high accuracy in identifying fake reviews, our system helps to preserve the reliability and trustworthiness of online review sites, empowering consumers to make more informed decisions.

## 6. FUTURE ENHANCEMENT

Enhancing the performance of existing models for fake review detection involves several key strategies. Model optimization, including hyper parameter tuning and other techniques, can significantly improve accuracy and efficiency. Fine-tuning parameters such as learning rates, tree depths, and the number of estimators can yield better performance. Additionally, feature engineering plays a crucial role in improving model effectiveness by incorporating metadata like user behavior patterns, review timestamps, and reviewer credibility. Exploring new techniques like advanced algorithms (e.g., GPT-3 or other transformer46 based models) and deep learning approaches (e.g., convolutional neural networks and recurrent neural networks) can provide deeper insights and a more nuanced understanding of text. Data, enhancing the model's robustness and accuracy. The developed methodologies can be applied to other domains such as spam detection, sentiment analysis, and opinion mining, adapting the hybrid approach to address various types of deceptive information across different platforms and industries.

## 7. REFERENCES

[1] Michael Andrew Hadiwijaya, Felix Putra Pirdaus, Darryl Andrews, Said Achmad, and Rhio Sutoyo. "Sentiment Analysis on Tokopedia Product Reviews sing Natural Language Processing". In: 2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS). IEEE. 2023, pp. 380–386.

[2] Pratiksha Shetgaonkar, Jane Trinity Rodrigues, Shailendra Aswale, Vialli Luis Keji Gonsalves, Jane Crystal Rodrigues, and Aditya Naik. "Fake review detection using sentiment analysis and deep learning". In: 2021 International Conference on Technological Advancements and Innovations (ICTAI). IEEE. 2021, pp. 140–145.

[3] VP Sumathi, SM Pudhiyavan, M Saran, and V Nandha Kumar. "Fake review detection of e-commerce electronic products using machine learning techniques". In: 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA). IEEE. 2021, pp. 1–5.

[4] Deepanshu Jain, Sayam Kumar, and Yashika Goyal. "Fake reviews filtering system using supervised machine learning". In: 2022 IEEE International Conference on Data Science and Information System (ICDSIS). IEEE. 2022, pp. 1–4.

[5] Digvijay Singh, Minakshi Memoria, and Rajiv Kumar. "Deep Learning Based Model for Fake Review Detection". In: 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT). IEEE. 2023, pp. 92–95.

[6] Mujahed Abdulqader, Abdallah Namoun, and Yazed Alsaawy. "Fake online reviews: A unified detection model using deception theories". In: IEEE Access 10 (2022), pp. 128622–128655.

[7] R Poonguzhali, S Franklin Sowmiya, P Surendar, and M Vasikaran. "Fake reviews detection using support vector machine". In: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). IEEE. 2022, pp. 1509–1512.

[8] Ting-You Lin, Basabi Chakraborty, and Chun-Cheng Peng. "A study on identification of important features for efficient detection of fake reviews". In: 2021 International Conference on Data Analytics for Business and Industry (ICDABI). IEEE. 2021, pp. 429–433.

[9] Syed Mohammed Anas and Santoshi Kumari. "Opinion mining based fake product review monitoring and removal system". In: 2021 6th International Conference on Inventive Computation Technologies (ICICT). IEEE. 2021, pp. 985–988.