Research Article BEU Fen Bilimleri Dergisi 15(1): 245-258 (2026) DOI: 10.17798/bitlisfen.1796956

BİTLİS EREN UNIVERSITY
FEN BİLİMLERİ DERGİSİ
ISSN: 2147-3129 / e-ISSN: 2147-3188

# AI vs. HUMAN TEXT DETECTION: A HIGH-ACCURACY ENSEMBLE APPROACH USING MACHINE LEARNING

**Yunus KÖKVER**

*Ankara University, Elmadağ Vocational School, Computer Technologies Department, Ankara, Türkiye, ykokver@ankara.edu.tr*

**Article Info**

***Received:*** *October 4, 2025*
***Revised:*** *December 12, 2025*
***Accepted:*** *February 9, 2026*

**Keywords**
*Natural Language Processing,*
*Artificial Intelligence and Ethics,*
*Machine Learning,*
*Ensemble Models,*
*Text Classification.*

**ABSTRACT**

This study aims to develop and evaluate a machine learning (ML)-based classification model for distinguishing between texts generated by artificial intelligence (AI) and those written by humans. Utilizing a comprehensive dataset comprising 487235 text samples, various ML algorithms—including Multilayer Perceptron (MLP), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), and an Ensemble Model—were trained and evaluated to classify AI-generated and human-generated texts. Ensemble Model, which combines the best-performing algorithms, achieved an accuracy rate of 99.90%, outperforming individual models. Additionally, the study presents a user-friendly interface that enables real-time classification of texts using the weights of the ensemble model. This interface holds potential as a practical tool for researchers and professionals in fields such as education, academia, and media. The model's generalization capability was also tested on a user-generated dataset through the user interface, and it was found to be consistent with the primary dataset, achieving an "Almost Perfect" level according to the Kappa statistic. This study highlights the necessity of robust tools to mitigate ethical and security risks associated with AI-generated content. Moreover, ensemble models show great promise in handling complex classification tasks.

## 1. INTRODUCTION

With rapid advances in natural language processing (NLP), the ability of AI to generate human-like text has reached unprecedented levels, bringing both opportunities and challenges. The emergence of advanced language models based on deep learning and transformer architectures has enabled AI-generated texts to closely resemble human writing in terms of fluency, coherence, and creativity [1]. This is particularly evident in models like GPT-3, where AI-generated responses are nearly indistinguishable from those produced by human authors [2]. However, this increasing realism has made it difficult to differentiate AI-generated content from human-created content in various domains such as education, journalism, and social media, raising significant ethical, security, and reliability concerns. To address this classification challenge, ML techniques and deep learning models have emerged as fundamental approaches in text classification [3].

Traditional ML classifiers, such as LR, SVM, and DT, have provided a fundamental approach for text classification. These methods have focused on handcrafted features, such as word frequency, n-grams, and syntactic features, to distinguish between content generated by humans and AI [4]. Such algorithms analyze the characteristic features of AI-generated texts, enabling the successful differentiation of their sources. When trained on large datasets, these models can capture fundamental linguistic differences between human-written and AI-generated content, making them effective for classification tasks [5]. These ML-based approaches, which rely on specific features rather than directly modeling attention mechanisms or contextual understanding, produce simpler yet interpretable results [6]. Studies aimed at separating AI-generated texts from original texts are discussed in detail under three subheadings in the following literature review section.

### 1.1. Literature Review

#### 1.1.1. Stylometry and Linguistic Analyses

Stylometry is a powerful method used to identify the author of a text and analyze its linguistic style. Research in this field has particularly focused on detecting stylistic changes in multi-author documents. Zamir et al. (2024) employed stylometric analysis techniques to identify author change points in multi-author documents [7]. Another study on how stylometric analysis can be used for author identification focused on analyzing stylometric features using ML techniques. These methods have provided reliable results in identifying authorship [8]. Especially on social media platforms, studies aimed at detecting AI-generated texts through stylometric features have increased. For example, Pascucci et al. (2020) used stylometry and ML methods to detect fake content in hotel reviews [9]. Another study combined stylometric analysis with artificial neural networks and fuzzy logic techniques for author identification [10]. In yet another study on stylometric analysis (2023), the focus was on detecting AI-generated texts in Twitter (X) timelines using stylometric features [11]. In the study conducted by Mikros et al. (2023), the detection of AI-generated texts was explored through the combination of stylometric features and transformer-based models [12].

#### 1.1.2. ML and Deep Learning Based Classification Methods

ML and deep learning-based classification methods play a significant role in detecting AI-generated texts. In the study conducted by Wang et al. [13], a 99.72% accuracy rate was achieved using the BERT algorithm for detecting AI-generated texts. In the study by Trandabat and Gifu [14], the importance of ML and RoBERTa classification methods in detecting AI-generated fake news was emphasized, and the potential application of these techniques on social media platforms was examined. Another study explored the successful differentiation between AI-generated and human-written texts using ML methods, and the success rates of the classifiers used were analyzed [15]. Alamleh et al. investigated the performance of various ML algorithms in distinguishing texts generated by ChatGPT [5]. In another study, the performance of various ML and deep learning-based detection tools was comprehensively compared [16]. Similarly, Gaggar et al. utilized deep learning and ML models, such as RoBERTa and SVM, to detect texts generated by large language models like GPT-3.5 [17]. In another study, a transformer-based model was proposed for the detection of AI-generated texts using a Large Language Models (LLM) based approach [18].

#### 1.1.3. Ethical and Safety Aspects

The capacity to discern texts generated by AI, along with the ethical and security ramifications of such technologies, has given rise to a series of substantial concerns, both from a technical standpoint and in terms of their societal implications. Zhang et al. proposed an innovative hybrid approach combining TF-IDF techniques and deep learning models to detect AI-generated texts. This study emphasizes the security dimension for more effective detection of fake content [19]. In a different study, the focus was on the potential of AI-generated texts to disseminate misinformation. This research provides solutions for detecting misinformation created by AI and suggests ways to prevent the spread of such content [20]. In the study conducted by Nguyen et al., the focus was on detecting AI-generated texts, and the ethical and security aspects of managing this process were discussed [21]. In another study highlighting the potential dangers of fake content production in scientific publications, it was emphasized that AI-generated fake scientific abstracts could lead to academic fraud, underscoring the need for robust tools to detect such content [22]. In another study focusing on plagiarism detection in AI-generated texts, it was discussed in detail that the lack of transparency in existing systems could pose ethical challenges [23].

Beyond plagiarism, the ethical implications of AI-generated content encompass the erosion of academic integrity and the potential for algorithmic hallucinations [24]. AI models can produce plausible but entirely fabricated information, which, if undetected, threatens the reliability of scientific literature and public discourse [25]. Furthermore, the lack of accountability in AI-generated texts complicates the traditional concept of authorship, as these models cannot take responsibility for the veracity or ethical consequences of the information they generate [25]. The inherent biases within training datasets also pose a risk of propagating stereotypes, making the development of high-accuracy detection tools a necessity for maintaining transparency and trust in the digital information ecosystem.

### 1.2. Objectives and Problem Statement

Based on the literature review and consideration of AI algorithms, the objectives of this study have been defined as follows:

The primary objective of this study is to facilitate the detection processes of AI-generated texts through the utilization of a user-friendly interface. In this context, a user interface has been developed using the weights of the classification algorithms and allows users to enter a text and get the classification result instantly. This interface is intended to serve as a practical tool for both academic researchers and professionals in fields such as media and education.

A further objective of the study is to evaluate the generalization capability of the models employed for the detection of AI-generated texts across a range of text types.

This study aims to provide both theoretical and practical contributions to the detection of AI-generated texts.

From this perspective, the problem statement of the study is to determine whether AI-generated texts of different types can be reliably detected using ML and the ensemble model created. The following sub-problems have been formulated to address this problem:

1. Which ML algorithm achieves the highest performance level in detecting AI-generated texts?

2. What is the performance level of the ensemble model in detecting AI-generated texts?

In light of these objectives and problem formulation, the novel contributions of this study can be summarized as follows:

This study differs from existing AI–text detection research in several key aspects. First, instead of relying on a single classifier, a heterogeneous ensemble architecture combining Multilayer Perceptron, Random Forest, and Logistic Regression is proposed, enabling complementary strengths of statistical learning and neural architectures to be jointly exploited. Second, unlike many previous studies that train models on limited datasets, the proposed approach is developed using a large-scale dataset comprising 111095 balanced and rigorously filtered samples, which significantly enhances robustness and generalization capacity. Third, beyond classification performance, this study operationalizes the model by embedding the trained ensemble weights into a user-friendly real-time detection interface, enabling practical applicability for academia, journalism, and education. Finally, the model's reliability is validated on an independent secondary dataset, demonstrating an "Almost Perfect" agreement level based on Cohen's Kappa, thereby evidencing strong external validity.

## 2. METHOD

This study adopted the widely recognized ADDIE instructional design model within a design-based research framework. The model comprises five fundamental phases: analysis, design, development, implementation and evaluation [26]. Figure 1 illustrates the design cycle and implementation process that was followed in this research.


*Figure 1. Design Cycle and implementation process of ADDIE [26], [27].*

### 2.1. ADDIE Design Model

The ADDIE model is often described as a linear framework; however, it is iterative and cyclical in nature, as evaluation is conducted at every stage [28].

This study adopted the ADDIE model because its structured, iterative and evaluation-centred framework aligns well with the sequential stages of data preparation, model training, system implementation and performance evaluation.

#### 2.1.1. Analysis

A review of the extant literature reveals that studies in this field predominantly focus on word and document analysis [13], [23]. Additionally, it is evident that the datasets used in these studies are often limited in size. The use of a dataset comprising 487235 samples in this study, the development of ensemble models for ML algorithms, the creation of a user interface, and the integration of trained AI model weights into this interface distinguish this work from existing literature. Consequently, this study serves as an exemplary model for other academics and researchers in the field.

#### 2.1.2. Design

In this stage of data mining processes, the Knowledge Discovery in Databases (KDD) process model was applied. In the extant literature, data mining process models such as SEMMA, CRISP-DM and KDD are generally examined. However, many researchers and data mining experts adopt the KDD model as it is more comprehensive and accurate [29]. Therefore, the KDD process model was preferred in this study, and its main phases are illustrated in Figure 2, which is adapted from the original KDD framework proposed in the literature.

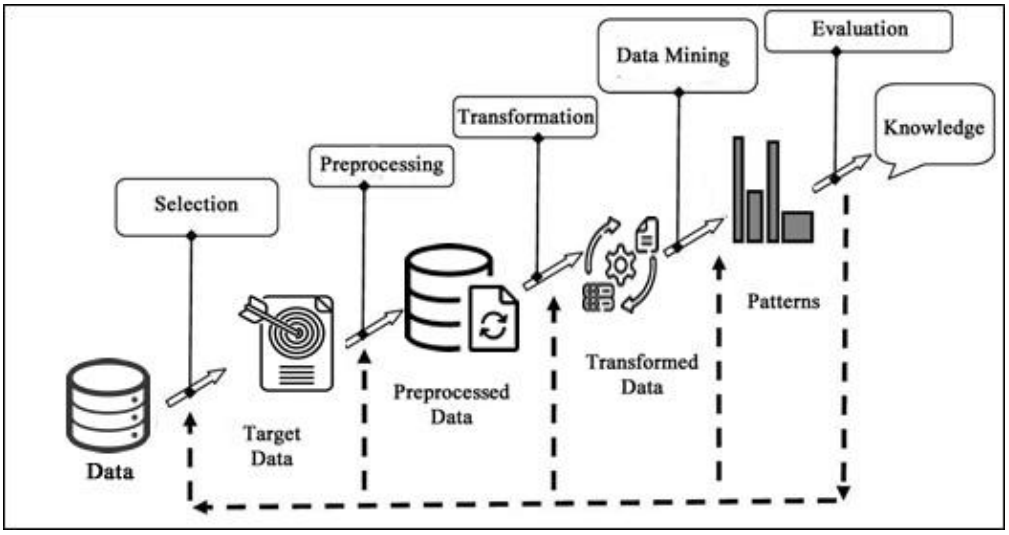


*Figure 2. Process model of KDD and its stages.*

The KDD model has an iterative and interactive structure. This model emerged from the need to analyze large-scale data. The KDD process is imperative for the identification of valid, novel, potentially useful, and comprehensible patterns within data, thereby facilitating knowledge discovery.

**Data:** The raw data of this study consists of the ai-vs-human-text [30] dataset consisting of 487235 data as the primary dataset.

**Data Selection:** At this stage, data set was checked and data with less than 20 sentences, data consisting of meaningless sentences and repetitive data were removed from the data set. In this way, 111095 data remained in the data set.

**Preprocessing:** To prepare plain text data for input into ML classification algorithms, the following preprocessing steps were applied:

Text Normalization: Converting the entire text to lowercase, removing punctuation marks from the text, eliminating line breaks in the text.

Tokenization: Splitting the text into word tokens, splitting the text into sentence tokens.

Stemming: Extracting the root forms of words.

Stopword Removal: Removing stopwords from the text.

Feature Extraction: Calculating root length distributions, determining word-based sentence length distributions, computing word richness ratios, calculating the average root length, determining the average sentence length in words, counting the total number of punctuation marks, counting the total number of stopwords used, counting the number of words written entirely in uppercase.

**Transformation:** Each text was preprocessed to extract its roots, and then feature vectors such as TF-IDF, bag-of-words vectors, and word-sentence distributions were generated. All resulting vectors were standardized using the min-max normalization method. As a result of these operations, the following features were obtained for each data point:

Type-Token Ratio (TTR): A measure of lexical diversity.

Average Word Length: The mean length of words in the text.

Average Sentence Length: The mean length of sentences in words.

Punctuation Count: The total number of punctuation marks.

Stopword Count: The total number of stopwords.

Uppercase Word Count: The number of words written entirely in uppercase.

Word Length Histogram Vector: A histogram vector based on letter counts for word lengths.

Sentence Length Histogram Vector: A histogram vector based on word counts for sentence lengths.

Bag-of-Words Vector: A vector representation of word frequencies.

TF-IDF Vector: A vector representation using Term Frequency-Inverse Document Frequency.

**Data Mining:** Following feature extraction, the dataset was divided into two parts: 80% for training and the remainder for testing. To mitigate overfitting and selection bias while ensuring that model performance is not influenced by the random selection of test data, the K-Folds Cross-Validation technique was applied. This validation method is commonly used to evaluate how well a model generalizes to independent datasets.

**Evaluation:** At this stage of the process, an objective analysis of the results is conducted using various evaluation metrics in order to assess the performance of the classification models. The metrics offer valuable insights into the model's accuracy and precision. In order to achieve this, four essential metrics derived from the Confusion Matrix – Accuracy, Recall, Precision and F1-Score – were utilised for assessment. Information on how to formulate these metrics is given in Table 1.

*Table 1. Formulation of evaluation criteria.*

| Evaluation Criteria | Formula |
|---|---|
| Accuracy | $\frac{TP + TN}{TP + FP + TN + FN}$ |
| Precision | $\frac{TP}{TP + FP}$ |
| Recall | $\frac{TP}{TP + FN}$ |
| F1-Score | $2 \text{ x } \frac{\text{Recall x Precision}}{\text{Recall} + \text{Precision}}$ |

***Note:*** *TP (true positives) denotes the number of positive samples that have been correctly classified, while TN (true negatives) signifies the number of negative samples that have been correctly classified. In contrast, FP (false positives) indicates negative samples that have been incorrectly classified as positive, and FN (false negatives) refers to positive samples that have been mistakenly classified as negative.*

### 2.1.3. Development

In this study, a primary dataset was used for training ML algorithms [30]. The data consists of two classes: '0' and '1'. The '0' class represents the 'Human Generated' category, while the '1' class corresponds to the 'AI Generated' category. The 'Human Generated' class contains 305797 samples, whereas the 'AI Generated' class includes 181438 samples. Statistical information, including the mean, minimum, and maximum values related to number of characters, number of words, and number of sentences, is presented in Table 2.

*Table 2. Statistical information for the primary dataset.*

| | Mean | Minimum | Maximum |
|---|---|---|---|
| **Number of Characters** | 2270.82 | 1 | 20373 |
| **Number of Words** | 439.49 | 1 | 4726 |
| **Number of Sentences** | 20.15 | 0 | 134 |

A secondary dataset was used to test the developed user interface. The data in the created secondary dataset consists of two categories: 'academic articles' and 'news texts'. The classification of these categorical data is labeled as '0: Human Generated' and '1: AI Generated'. During the Data Selection phase, samples with fewer than 20 sentences were removed from the dataset, leaving a total of 1226 samples. Of these, 675 belong to the 'Human Generated' category, and 551 belong to the 'AI Generated' category.

The 'Human Generated' category consists of texts from the Introduction sections of academic papers published before 2019, as well as from news outlets such as the New York Times, Reuters, and Associated Press. The reason for selecting academic papers published before 2019 is that large language models began to resemble human-generated texts with the introduction of GPT-2 in 2019 [31]. The data in the 'AI Generated' category consists of samples from ChatGPT's GPT-4 version (276 samples) and the Google Gemini 1.5 Flash version (275 samples). Statistical information regarding the secondary dataset is provided in Table 3.

*Table 3. Statistical information for the secondary dataset.*

| | Mean | Minimum | Maximum |
|---|---|---|---|
| **Number of Characters** | 1865.43 | 1367 | 16686 |
| **Number of Words** | 185.78 | 128 | 3119 |
| **Number of Sentences** | 23.15 | 20 | 102 |

### 2.1.4. Implementation

At this stage of the study, a user interface was created using the weights of the ensemble model. The joblib library, which is part of the sklearn library in Python, was used to save the model weights. These weights were saved in a (.pkl) file format and utilized in the interface. The user interface is shown in Figure 3.

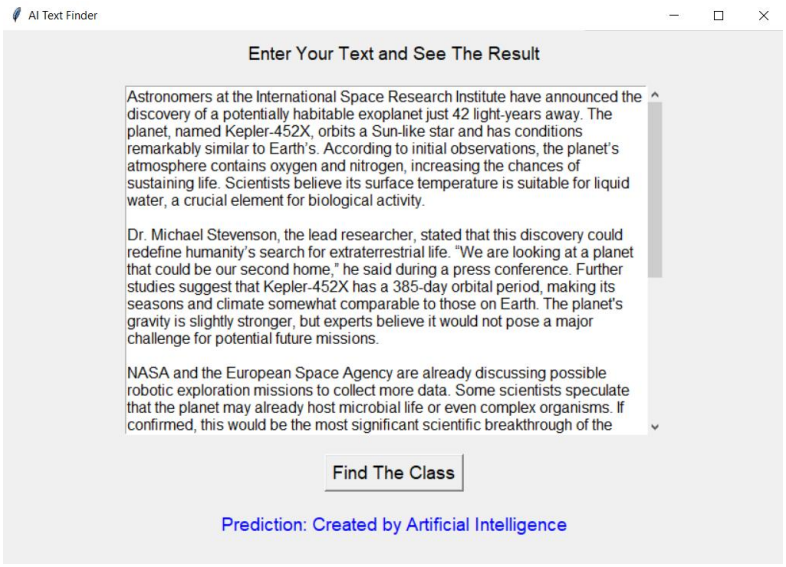


*Figure 3. User Interface.*

### 2.1.5. Evaluation

At this stage, the data in the secondary dataset were classified using the user interface. To calculate the agreement level between the classification performance of the primary and secondary datasets, the Kappa (κ) statistic was used. The κ statistic is designed to measure the level of agreement between two raters in classification tasks. The application of the κ statistic is predicated on specific assumptions, as outlined by Brennen and Prediger (1981), namely that the objects or individuals being classified must be independent, the raters' evaluations should not influence each other, and the categories used for scoring must also be independent.

Cohen's Kappa is computed using the standard formula;

$$\kappa = \frac{P_{observed} - P_{expected}}{1 - P_{expected}} \tag{1}$$

In the interpretation of the κ statistic, the agreement levels are presented in Table 4.

*Table 4. Value ranges for the interpretation of the Kappa statistic.*

| κ | Strength of Agreement |
|---|---|
| <0.00 | Poor |
| 0.00-0.20 | Slight |
| 0.21-0.40 | Fair |
| 0.41-0.60 | Moderate |
| 0.61-0.80 | Substantial |
| 0.81-1.00 | Almost Perfect |

## 2.2. ML Algorithms Used in the Study

The ML algorithms used in this study are detailed below.

**2.2.1. Multilayer Perceptron (MLP):** MLP is a fundamental type of artificial neural network that processes input data through multiple layers. This renders it an effective method for addressing nonlinear problems. The MLP comprises an input layer, one or more hidden layers, and an output layer. In this structure, each neuron receives data from connected neurons, multiplies it by weights, adds a bias, and applies an activation function. The training process is typically conducted using the backpropagation algorithm along with derivative-based optimization techniques.

**2.2.2. Support Vector Machine (SVM):** SVM is a frequently employed supervised learning algorithm for both classification and regression tasks. The primary objective of an SVM is to ascertain the optimal hyperplane that most effectively differentiates data points across various classes. In instances where the data is not linearly separable, an approach known as the "kernel trick" is employed by the SVM to map the data into a higher-dimensional space, thereby facilitating enhanced separation. Additionally, it improves the model's generalization performance by penalizing misclassified data points through a regularization parameter (C).

**2.2.3. Gradient Boosting (GB):** GB is a supervised ML technique that constructs powerful predictive models by sequentially combining multiple weak learners, most commonly decision trees. It improves overall prediction accuracy by allowing each new model to learn from the mistakes of its predecessors. The error correction process is carried out by calculating the negative gradient of the loss function relative to the existing model. Gradient Boosting is extensively applied in both classification and regression tasks due to its high accuracy. Moreover, it incorporates regularization methods to reduce the risk of overfitting.

**2.2.4. Decision Tree (DT):** DT is a supervised ML algorithm that is widely used for both classification and regression problems due to its simplicity and effectiveness. DTs classify or predict outcomes by recursively splitting the dataset into branches based on a series of 'decisions'. The model typically selects the best splitting feature using metrics such as information gain, Gini index, or variance reduction. Starting from the root node, the data is divided at each node based on specific feature values until leaf nodes are reached. Due to its interpretable structure, decision trees are ideal for exploratory data analysis and problems requiring explainability. However, to prevent overfitting, pruning techniques or depth constraints are often applied.

**2.2.5. Random Forest (RF):** RF is a robust and versatile supervised ML algorithm that is commonly applied to classification and regression tasks. It builds an ensemble model by training multiple decision trees on different subsets of the dataset, with each tree using a randomly chosen subset of features. The final prediction is obtained by majority voting for classification or by averaging the outputs for regression. By leveraging random sampling of both data and features, RF minimises overfitting and enhances the model's ability to generalise effectively. In addition to providing high levels of accuracy, RF is effective in handling missing data and complex relationships within datasets, as well as offering advantages in assessing feature importance and enhancing model interpretability.

**2.2.6. Logistic Regression (LR):** LR is a popular supervised ML algorithm designed for classification tasks. The algorithm attempts to classify data using a linear model and expresses classification outcomes as probabilities. LR applies the sigmoid activation function to transform predicted values into the range [0,1]. The primary goal of the model is to maximize the log-likelihood function to optimize classification. Model parameters are typically optimized using gradient descent or its variants. Due to its simplicity and effectiveness, LR is a fundamental approach for binary classification problems and is commonly used in various fields such as credit risk assessment, medical diagnosis, and marketing analytics.

**2.2.7. Ensemble Model Architecture**

The Ensemble Model Architecture is an approach that combines multiple ML algorithms to enhance overall prediction accuracy while mitigating the limitations of individual models. In this architecture, the predictions of various base models (weak learners) are aggregated, with the final decision typically made using voting, weighted averaging, or other combination strategies. Ensemble methods are generally categorized into three types: Bagging, Boosting, and Stacking. Bagging trains models independently on different subsets of the data; Boosting focuses on sequentially improving model performance by correcting the errors of the previous models; and Stacking integrates the outputs of different algorithms using a meta-learner. Ensemble models are widely used in classification and regression problems requiring high accuracy and robust performance in complex data structures. This classifier is often observed to be more accurate than any of the individual classifiers forming the ensemble.

The comprehensive workflow of the proposed model is presented in Figure 4.

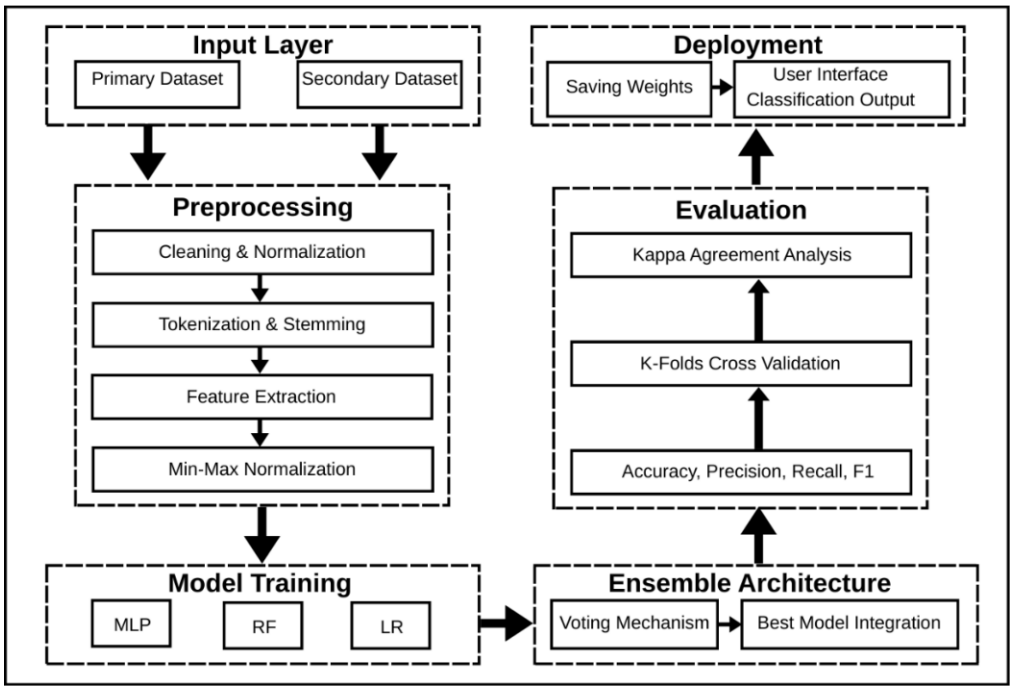


*Figure 4. Block diagram of the proposed system.*

## 3. FINDINGS

### 3.1. RQ#1. Which ML algorithm achieves the highest performance level in detecting AI-generated texts?

When examining Table 5, it is observed that the ML classifier with the highest validation accuracy is MLP.

*Table 5. ML Classifiers and evaluation metrics.*

| ML Classifier | Class | Recall | Precision | F1-Score | Accuracy (%) |
|---|---|---|---|---|---|
| MLP | 0 | 0.9997 | 0.9990 | 0.9993 | 99.89 |
| | 1 | 0.9955 | 0.9985 | 0.9970 | |
| RF | 0 | 0.9999 | 0.9951 | 0.9975 | 99.59 |
| | 1 | 0.9779 | 0.9997 | 0.9887 | |
| LR | 0 | 0.9999 | 0.9921 | 0.9960 | 99.34 |
| | 1 | 0.9640 | 0.9997 | 0.9816 | |
| DT | 0 | 0.9955 | 0.9922 | 0.9939 | 98.99 |
| | 1 | 0.9645 | 0.9796 | 0.9720 | |
| GB | 0 | 0.9938 | 0.9902 | 0.9920 | 98.69 |
| | 1 | 0.9556 | 0.9715 | 0.9635 | |
| SVM | 0 | 0.9949 | 0.9834 | 0.9891 | 98.20 |
| | 1 | 0.9241 | 0.9756 | 0.9491 | |

0: Human Generated, 1: AI Generated

The MLP classifier achieved the highest performance with 99.89% accuracy, reaching 0.9997 recall and 0.9990 precision for human-generated texts. For AI-generated texts, it obtained 0.9955 recall and 0.9985 precision. RF classifier ranked second with 99.59% accuracy, followed by LR with 99.34% accuracy. DT, GB, and SVM classifiers demonstrated balanced performances with accuracy rates of 98.99%, 98.69%, and 98.20%, respectively.

Although accuracy is an essential metric for evaluating model performance, it can be misleading, especially in imbalanced datasets. Therefore, recall, precision, and their harmonic mean, F1-score, are

crucial for a more in-depth analysis of the model's effectiveness. Low recall means the model misses many positive samples, which can be critical in applications like cancer diagnosis. Low precision indicates high false positive rates, leading to unnecessary alarms in decision-making systems. F1-score provides a balanced assessment of the model's performance by considering both recall and precision. According to Table 5, the recall, precision, and F1-score values are consistent with accuracy for all algorithms, indicating that the dataset is well-balanced across classes. The confusion matrix details for each ML algorithm are provided in Appendix 1.

### 3.2. RQ#2. What is the performance level of the ensemble model in detecting AI-generated texts?

Among the top three classifiers (MLP, RF, and LR), the Ensemble Model, developed using a voting mechanism, achieved the highest accuracy of 99.90%, as shown in Table 6. This result confirms that the ensemble approach effectively classifies both human- and AI-generated texts with high accuracy. Additionally, the model demonstrated high reliability in classification, as indicated by the Recall, Precision, and F1-score metrics. The confusion matrix details for the Ensemble Model are provided in Appendix 1.

*Table 6. Evaluation metrics of the Ensemble Model.*

| Classifier | Class | Recall | Precision | F1-Score | Accuracy (%) |
|---|---|---|---|---|---|
| **Ensemble Model** | 0 | 0.9997 | 0.9991 | 0.9994 | 99.90 |
| | 1 | 0.9960 | 0.9985 | 0.9973 | |

### 3.3. RQ#3. Is there a significant difference between the performance levels obtained from the primary and secondary datasets used in the study?

In this study, the κ statistic was used to assess the agreement between the ensemble model classification (Ensemble), trained using the primary dataset, and the classification performed using the secondary dataset (Human). The κ statistic was calculated using SPSS, and the Crosstabulation and Symmetric Measures results are presented in Table 7 and Table 8, respectively. Here, "Human" represents the data labeled by the human expert (secondary dataset), while "Ensemble" refers to the results produced by the user interface developed using the ensemble model trained on the primary dataset.

*Table 7. Human-Ensemble Cross-tabulation.*

| Statistics Count | | Ensemble | | Total |
|---|---|---|---|---|
| | | Human Generated | AI Generated | |
| **Human** | **Human Generated** | 586 | 27 | 613 |
| | **AI Generated** | 89 | 524 | 613 |
| **Total** | | 675 | 551 | 1226 |

*Table 8. Symmetric measurement results.*

| | | Value | Asymptotic Standard Error[a] | Approximate T[b] | Approximate Significance (p) |
|---|---|---|---|---|---|
| **Measure of Agreement** | **Kappa** | 0.811 | 0.017 | 28.535 | 0.000* |
| **N of Valid Cases** | | 1226 | | | |

*p<0.05**

*a. Not assuming the null hypothesis.*

*b. Using the asymptotic standard error assuming the null hypothesis.*

Upon examining Table 8, it is observed that the κ value is 0.811, indicating an 'Almost Perfect' level of agreement according to the criteria in Table 4. Additionally, the Approximate Significance value (p <

0.05) suggests that there is a statistically significant difference between the performance levels obtained from the primary dataset and the secondary dataset used in this study.

## 4. DISCUSSION

Previous studies in the field of AI-generated text detection, although successful, present several notable limitations. A considerable number of existing works rely on relatively small or private datasets, which restricts the generalisation capability and reproducibility of their findings. In addition, many of these studies depend on a single classifier or model architecture, making their performance highly sensitive to dataset characteristics. Furthermore, external validation through an independent dataset is rarely reported, meaning that real-world robustness remains insufficiently demonstrated. To address these shortcomings, the present study employs a large-scale publicly accessible dataset, adopts a heterogeneous ensemble model that combines the complementary strengths of different machine learning algorithms, and validates its reliability using a secondary dataset, demonstrating "Almost Perfect" agreement based on Cohen's Kappa. These aspects collectively strengthen the robustness, reproducibility, and practical applicability of the proposed approach compared to previous studies.

As illustrated in Table 9, a comparative analysis of the results obtained from the proposed method in this study is presented, considering the size of the dataset used, along with the performance of state-of-the-art models from the literature.

*Table 9. A comparative analysis of the results obtained from the proposed method and other methods in the literature.*

| Study | Dataset Source | Dataset Size | Classifier | Accuracy (%) |
|---|---|---|---|---|
| Wang et al.[13] | Private | 1378 | BERT | 99.72 |
| Mo et al. [18] | Private | 1378 | LSTM+CNN | 99 |
| Martinelli et al. [32] | AI-human-text [33] | 10000 | albert-large-v2 | 96 |
| Alamleh et al. [5] | Private | 500 | RF | 93.50 |
| Trandabat&Gifu [14] | LIAR [34] + Kdnuggets [35] + Private | 20000 | RoBERTa | 89.3 |
| **Proposed Method** | **ai-vs-human-text** [30] | **111095** | **Ensemble Model** | **99.9** |

The comparative analysis presented in Table 9 demonstrates that this study, with an accuracy rate of 99.90%, is among the highest-performing studies in the existing literature. Specifically, the BERT-based model proposed by Wang et al. [13] achieved an accuracy of 99.72%, while the LSTM+CNN model developed by Mo et al. [18] attained an accuracy of 99%. In this study, an accuracy rate of 99.90% was achieved using an ensemble model, surpassing the performance of other studies in the literature. This result highlights that ensemble models can outperform individual models and that integrating multiple ML algorithms can enhance classification performance.

One of the most remarkable aspects of this study, compared to other studies in the literature, is the size and diversity of the dataset used. For instance, Wang et al. [13] utilized a dataset containing 1378 samples, Mo et al. [18] also employed 1378 samples, and Martinelli et al. [32] used a dataset with 10000 samples. In contrast, this study leveraged a significantly larger dataset comprising 111095 samples, enhancing the model's generalization capability. This extensive dataset has enabled the model to achieve high accuracy across various text types. Notably, when compared to the model trained on the 10000-sample dataset used by Martinelli et al. [32] , which achieved an accuracy rate of 96%, the proposed approach demonstrates a substantial improvement, reaching an accuracy of 99.90%.

Among the ML algorithms utilized in this study, MLP demonstrated the highest performance, achieving an accuracy of 99.89%. Additionally, other algorithms such as RF and LR also produced highly successful results, with accuracy rates of 99.59% and 99.34%, respectively. However, the ensemble model, which combines these individual algorithms, achieved an even higher accuracy of 99.90%, demonstrated superior performance compared to standalone models. Furthermore, when compared to

transformer-based deep learning models used in previous studies, the ensemble model employed in this study attained a higher accuracy, highlighting its effectiveness in classification tasks.

The high accuracy rate achieved by the proposed ensemble model carries significant ethical importance. In an era where AI-generated misinformation can be rapidly disseminated, a highly reliable detection system serves as a safeguard for information integrity. By minimizing false positives and negatives, this study provides a robust tool for educators and media professionals to uphold ethical standards and prevent the misuse of AI in high-stakes domains such as academic publishing and journalism.

## 5. CONCLUSION

This study presents an effective classification model for distinguishing AI-generated texts from human-written texts. With the rapid advancement of AI technologies and the widespread adoption of large language models, the realism of AI-generated texts has significantly increased, making their differentiation from human-written texts increasingly challenging. This challenge raises ethical, security, and reliability concerns, particularly in fields such as education, journalism, and academic publishing. To address these issues, this study evaluates various ML algorithms and an ensemble model for detecting AI-generated texts. The results indicate that ensemble models can outperform individual classifiers in complex classification tasks.

Additionally, the user interface developed in this study makes the detection of AI-generated texts practical and accessible. By leveraging the weights of the trained ensemble model, the interface enables the rapid classification of texts. Furthermore, the $\kappa$ statistical analysis demonstrates an "Almost Perfect" level of agreement between human experts and the AI system, confirming the model's reliability for academic articles and news texts.

One limitation of this study is that the dataset consists exclusively of English texts, preventing an assessment of classification performance for texts in other languages. Future research could explore the model's performance across different languages and text types, as well as evaluate its effectiveness on more extensive and diverse datasets.

**Statement of Research and Publication Ethics**

The study is complied with research and publication ethics.

**Artificial Intelligence (AI) Contribution Statement**

AI tools were used only to assist with English language translation and editing. The scientific content, data analysis, results, and conclusions were fully developed by the authors without AI assistance.

## REFERENCES

[1] Y. Kökver, H. M. Pektaş, and H. Çelik, "Artificial intelligence applications in education: Natural language processing in detecting misconceptions," *Educ. Inf. Technol.*, pp. 1–32, Aug. 2024, doi: 10.1007/S10639-024-12919-1/FIGURES/2.

[2] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, May 2020, Accessed: Oct. 17, 2024. [Online]. Available: https://arxiv.org/abs/2005.14165v4

[3] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. 2019 Conf. North*, pp. 4171–4186, 2019, doi: 10.18653/V1/N19-1423.

[4] C. D. Manning, "Introduction to information retrieval," 2008, *Cambridge university press*.

[5] H. Alamleh, A. A. S. Alqahtani, and A. Elsaid, "Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning," *2023 Syst. Inf. Eng. Des. Symp. SIEDS 2023*, pp. 154–158, 2023, doi: 10.1109/SIEDS58326.2023.10137767.

[6] M. Nour, B. Arabacı, H. Öcal, and K. Polat, "New approaches to epileptic seizure prediction based on EEG signals using hybrid CNNs," *Int. J. Intell. Eng. Informatics*, vol. 12, no. 1, pp. 85–102, 2024, doi: 10.1504/IJIEI.2024.137706.

[7] M. T. Zamir, M. A. Ayub, A. Gul, N. Ahmad, and K. Ahmad, "Stylometry Analysis of Multi-authored Documents for Authorship and Author Style Change Detection," Jan. 2024, Accessed: Oct. 15, 2024. [Online]. Available: https://arxiv.org/abs/2401.06752v1

[8] A. de Pablo, O. Araque, and C. A. Iglesias, "Radical Text Detection based on Stylometry," in *International Conference on Information Systems Security and Privacy*, Science and Technology Publications, Lda, 2020, pp. 524–531. doi: 10.5220/0008971205240531.

[9] A. Pascucci, R. Manna, C. Caterino, V. Masucci, and J. Monti, "Is this hotel review truthful or deceptive? A platform for disinformation detection through computational stylometry," in *Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management*, 2020, pp. 35–40.

[10] H. El-Fiqi, E. Petraki, and H. A. Abbass, "A computational linguistic approach for the identification of translator stylometry using Arabic-English text," *IEEE Int. Conf. Fuzzy Syst.*, pp. 2039–2045, 2011, doi: 10.1109/FUZZY.2011.6007535.

[11] T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, and H. Liu, "Stylometric detection of ai-generated text in twitter timelines," *arXiv Prepr. arXiv2303.03697*, 2023.

[12] G. K. Mikros, A. Koursaris, D. Bilianos, and G. Markopoulos, "AI-Writing Detection Using an Ensemble of Transformers and Stylometric Features.," in *IberLEF@ SEPLN*, 2023.

[13] H. Wang, J. Li, and Z. Li, "AI-generated text detection and classification based on BERT deep learning algorithm," *Theor. Nat. Sci.*, vol. 39, no. 1, pp. 312–317, Jul. 2024, doi: 10.54254/2753-8818/39/20240625.

[14] D. Trandabat and D. Gifu, "Discriminating AI-generated Fake News," *Procedia Comput. Sci.*, vol. 225, pp. 3822–3831, Jan. 2023, doi: 10.1016/J.PROCS.2023.10.378.

[15] R. Kumar and M. Mindzak, "Who wrote this? Detecting artificial intelligence–generated text from human-written text," *Can. Perspect. Acad. Integr.*, vol. 7, no. 1, 2024.

[16] A. Akram, "An empirical study of ai generated text detection tools," *arXiv Prepr. arXiv2310.01423*, 2023, doi: https://doi.org/10.48550/arXiv.2310.01423.

[17] R. Gaggar, A. Bhagchandani, and H. Oza, "Machine-generated text detection using deep learning," *arXiv Prepr. arXiv2311.15425*, 2023.

[18] Y. Mo, H. Qin, Y. Dong, Z. Zhu, and Z. Li, "Large Language Model (LLM) AI text generation detection based on transformer deep learning algorithm," Apr. 2024, doi: 10.48550/arxiv.2405.06652.

[19] Y. Zhang *et al.*, "Enhancing Text Authenticity: A Novel Hybrid Approach for AI-Generated Text Detection," Jun. 2024, Accessed: Oct. 15, 2024. [Online]. Available: https://arxiv.org/abs/2406.06558v1

[20] A. Najee-Ullah, L. Landeros, Y. Balytskyi, and S. Y. Chang, "Towards Detection of AI-Generated Texts and Misinformation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13176 LNCS, pp. 194–205, 2022, doi: 10.1007/978-3-031-10183-0_10/FIGURES/4.

[21] T. T. Nguyen, A. Hatua, and A. H. Sung, "How to Detect AI-Generated Texts?," *2023 IEEE 14th Annu. Ubiquitous Comput. Electron. Mob. Commun. Conf. UEMCON 2023*, pp. 464–471, 2023, doi: 10.1109/UEMCON59035.2023.10316132.

[22] P. C. Theocharopoulos, P. Anagnostou, A. Tsoukala, S. V. Georgakopoulos, S. K. Tasoulis, and V. P. Plagianakos, "Detection of Fake Generated Scientific Abstracts," *Proc. - IEEE 9th Int. Conf. Big Data Comput. Serv. Appl. BigDataService 2023*, pp. 33–39, 2023, doi: 10.1109/BIGDATASERVICE58306.2023.00011.

[23] M. A. Quidwai, C. Li, and P. Dube, "Beyond black box ai-generated plagiarism detection: From sentence to document level," *arXiv Prepr. arXiv2306.08122*, 2023.

[24] Y. Zhang, T. Zhou, H. Qiao, and T. Li, "Ethical Issues in AI-Generated Texts: A Systematic Review and Analysis," *Int. J. Human–Computer Interact.*, pp. 1–28, 2025.

[25] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, 2023.

[26] R. M. Branch, "Instructional design: The ADDIE approach," *Instr. Des. ADDIE Approach*, pp. 1–203, 2010, doi: 10.1007/978-0-387-09506-6/COVER.

[27] O. Karamustafaoğlu and H. M. Pektaş, "Developing students' creative problem solving skills with inquiry-based STEM activity in an out-of-school learning environment," *Educ. Inf. Technol.*, vol. 28, no. 6, pp. 7651–7669, Jun. 2023, doi: 10.1007/S10639-022-11496-5/TABLES/4.

[28] T. Trust and E. Pektas, "Using the ADDIE Model and Universal Design for Learning Principles to Develop an Open Online Course for Teacher Professional Development," *J. Digit. Learn. Teach. Educ.*, vol. 34, no. 4, pp. 219–233, Oct. 2018, doi: 10.1080/21532974.2018.1494521.

[29] U. Shafique and H. Qaiser, "A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)," *Int. J. Innov. Sci. Res.*, vol. 12, no. 1, pp. 217–222, 2014.

[30] S. Gerami, "AI Vs Human Text." Accessed: Oct. 17, 2024. [Online]. Available:

https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text

[31] "Better language models and their implications | OpenAI." Accessed: Dec. 22, 2024. [Online]. Available: https://openai.com/index/better-language-models/

[32] F. Martinelli, F. Mercaldo, L. Petrillo, and A. Santone, "A Method for AI-generated sentence detection through Large Language Models," *Procedia Comput. Sci.*, vol. 246, no. C, pp. 4853–4862, Jan. 2024, doi: 10.1016/J.PROCS.2024.09.351.

[33] "andythetechnerd03/AI-human-text · Datasets at Hugging Face." Accessed: Jan. 03, 2025. [Online]. Available: https://huggingface.co/datasets/andythetechnerd03/AI-human-text

[34] W. Y. Wang, "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection," *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, pp. 422–426, 2017, doi: 10.18653/V1/P17-2067.

[35] "Data Science, Machine Learning, AI & Analytics - KDnuggets." Accessed: Jan. 03, 2025. [Online]. Available: https://www.kdnuggets.com/