

# Whodunit: Classifying Code as Human Authored or GPT-4 generated- A case study on CodeChef problems

Oseremen Joy Idialu  
j2idialu@uwaterloo.ca  
University of Waterloo  
Waterloo, Ontario, Canada

Noble Saji Mathews  
noblesaji.mathews@uwaterloo.ca  
University of Waterloo  
Waterloo, Ontario, Canada

Rungroj Maipradit  
rungroj.maipradit@uwaterloo.ca  
University of Waterloo  
Waterloo, Ontario, Canada

Joanne M. Atlee  
jmatlee@uwaterloo.ca  
University of Waterloo  
Waterloo, Ontario, Canada

Meiyappan Nagappan  
mei.nagappan@uwaterloo.ca  
University of Waterloo  
Waterloo, Ontario, Canada

## ABSTRACT

Artificial intelligence (AI) assistants such as GitHub Copilot and ChatGPT, built on large language models like GPT-4, are revolutionizing how programming tasks are performed, raising questions about whether code is authored by generative AI models. Such questions are of particular interest to educators, who worry that these tools enable a new form of academic dishonesty, in which students submit AI-generated code as their work. Our research explores the viability of using code stylometry and machine learning to distinguish between GPT-4 generated and human-authored code. Our dataset comprises human-authored solutions from CodeChef and AI-authored solutions generated by GPT-4. Our classifier outperforms baselines, with an F1-score and AUC-ROC score of 0.91. A variant of our classifier that excludes gameable features (e.g., empty lines, whitespace) still performs well with an F1-score and AUC-ROC score of 0.89. We also evaluated our classifier with respect to the difficulty of the programming problem and found that there was almost no difference between easier and intermediate problems, and the classifier performed only slightly worse on harder problems. Our study shows that code stylometry is a promising approach for distinguishing between GPT-4 generated code and human-authored code.

## KEYWORDS

code stylometry, chatgpt, AI code, GPT-4 generated code, authorship profiling, software engineering

## 1 INTRODUCTION

AI tools like Github Copilot [24], ChatGPT [42], and Code Whisperer [46] are disrupting how educators teach and assess programming. These tools are promoted as coding “assistants” that aim to improve developer productivity by suggesting code snippets, bug fixes, code refactorings, and test cases. The use of AI assistants in introductory programming courses has been shown to increase the productivity of novice programmers in solving introductory programming problems, both in terms of improving the quality of their programs and easing the cognitive load and development effort required [32].

Eventually, however, educators need to assess how well students can program *without* the aid of their coding assistants, and they are worried about academic dishonesty. Programming courses already suffer from high levels of plagiarism [5] and contract cheating [13].

The ease with which these new tools can automatically generate code raises concerns about a new form of academic dishonesty, in which students submit AI-generated programs as their work [7, 22, 36]. Existing approaches to detecting plagiarism among student-submitted programs rely on automated similarity comparison tools, but these tools are unlikely to detect AI-generated solutions because AI-generated code has low similarity to student-authored code [45].

The goal of our research is to determine the viability of constructing a classifier that can distinguish between GPT-4 generated and human-authored code. We hypothesize that the low similarity between student-authored and GPT-4 generated code suggests that code stylometry and machine-learning classification can be used to distinguish between the two.

In this paper we make the following contributions: (1) We make a best-effort attempt at constructing a classifier for detecting GPT-4 generated Python code, using a combination of supervised machine learning (XGBoost [12]), and a collection of 140 code-stylometry features. The classifier is trained and evaluated on a dataset comprising 798 human-authored solutions and 798 GPT-4 generated solutions to 399 Python problems of varying degrees of difficulty from CodeChef.<sup>1</sup> To our knowledge, this is the first attempt to construct a classifier for detecting GPT-4 generated code based on a dataset with over 1000 solutions. (2) Our evaluation focuses on four research questions:

### RQ1: How well can code-stylometry features distinguish human-authored code from GPT-4 generated code?

Prior work has shown that a classifier trained on code-stylometry features can distinguish among different human developers. We conjecture that a classifier can be built to differentiate code developed by humans from code generated by AI tools.

### RQ2: How influential are non-gameable features in differentiating human-authored vs. GPT-4 generated code?

Coding styles are deemed *gameable* if they can be easily and strategically altered or avoided with minimal effort, particularly to mask the AI-generated nature of the code. Examples of such gameable styles include the use of empty lines and whitespace for readability, which can be quickly adjusted without significantly impacting the overall code structure or functionality. If a classifier relies heavily on gameable features, then it may be relatively easy to disguise a GPT-4 generated solution through

<sup>1</sup><https://www.codechef.com/>

simple code edits. We hypothesize that a classifier built using non-gameable code-stylometry features can effectively identify GPT-4 generated code.

### RQ3: How well does the classifier perform when trained and evaluated on only correct solutions?

We hypothesize that a dataset containing incorrect GPT-4 generated solutions may exhibit distinctive characteristics that could potentially enhance a classifier’s ability to identify GPT-4 generated solutions.

### RQ4: How well does the classifier perform when trained and evaluated across varying levels of problem difficulty?

We hypothesize that as the complexity of coding problems increases, the discriminative features between human and GPT-4 generated code may become more pronounced. This could be attributed to the unique problem-solving approaches employed by human developers compared to AI systems when faced with complex programming tasks.

(3) Our dataset comprising 399 problems and 1596 human and GPT-4 generated solutions is itself a contribution that we make publicly available to other researchers working on the problem of identifying GPT-4 generated code. We also provide a subset of the dataset comprising 161 problems whose solutions have been checked for correctness.

We provide a replication package,<sup>2</sup> which includes raw data, feature lists, and code scripts.

## 2 MOTIVATION AND RELATED WORK

In this section, we provide a real-world example of code-style differences between human and GPT-4 generated code to motivate our study and related work.

### 2.1 Motivation

Figure 1 presents two Python programs, both of which compute the sum of all palindromic numbers within an input range of integers<sup>3</sup> Figure 1a represents a user submission obtained from CodeChef, whereas Figure 1b was generated by ChatGPT. The two solutions exhibit clear differences in coding styles. The AI-generated code includes empty lines, and helper functions, whereas the human-authored code uses shorter identifiers. These differences and other code-style patterns (e.g., frequency of different keywords, complexity of expressions) used in prior studies on author attribution led us to question whether we could use code-stylometry features to build a classifier that distinguishes between human-authored code and GPT-4 generated code.

### 2.2 Related Work

**2.2.1 Detecting AI-generated Code.** The prevalence and potency of AI assistants have led researchers to start investigating the problem of detecting code generated by AI assistants. Puryear and Sprint [45] investigated how well-established plagiarism detection tools, MOSS [4], Codequiry [1], and CopyLeaks [2], could detect Copilot-generated solutions among a set of data science programming assignments. They found that Copilot-generated solutions

<sup>2</sup><https://zenodo.org/doi/10.5281/zenodo.10152237>

<sup>3</sup><https://www.codechef.com/problems/SPALNUM>

```

1 for _ in range(int(input())):
2     l, r = map(int, input().split())
3     result = 0
4     for i in range(1, r + 1):
5         if str(i) == str(i)[::-1]:
6             result += i
7     print(result)

```

(a) Example of human code

```

1 def is_palindrome(n):
2     return str(n) == str(n)[::-1]
3
4 def palindromic_numbers_sum(l, r):
5     total = 0
6     for n in range(1, r+1):
7         if is_palindrome(n):
8             total += n
9     return total
10
11 t = int(input())
12
13 for i in range(t):
14     l, r = map(int, input().split())
15     result = palindromic_numbers_sum(l, r)
16     print(result)

```

(b) Example of ChatGPT code

Figure 1: Solutions to a palindrome problem from CodeChef

exhibited little similarity to solutions authored by students. The highest observed similarity, identified by MOSS at 36%, fell well below the thresholds of similarity between student solutions that suggest plagiarism. Moreover, when “similar” Copilot and student solutions were manually inspected, the researchers determined that code similarities often reflected standard, commonly employed coding solutions or expected variable declarations. In work that is closest to ours, Bukhari et al. [10] attempt to use machine learning to distinguish between 28 student-authored and 30 AI-generated solutions for a C-language programming assignment involving singly-linked lists. Their approach leverages lexical and syntactic features in conjunction with multiple machine-learning models, achieving an accuracy rate of 92%.

We are also starting to see commercial tools such as Hacker-Rank [26] and Coderbyte [14] that claim to identify AI-generated code within user-submitted code. Unfortunately, evidence of their performance has not been provided and is not freely available for third-party evaluation.

Our study expands on the body of work in this emerging field, employing a more diverse problem set and more descriptive features for interpretability compared to the study by Bukhari et al. [10].

**2.2.2 Code Stylometry.** A related research problem focuses on identifying code authorship, typically by using code stylometry, which analyzes distinct coding styles that reflect patterns in the way programmers write code. There exists a substantial body of work on coding constructs that can serve as distinctive identifiers of individual coding styles. Pioneering work by Oman and Cook in 1989 [41] analyzed the authorship of 18 distinct Pascal programs published in six independently authored computer science textbooks. More recent studies have used code stylometry for authorship attribution

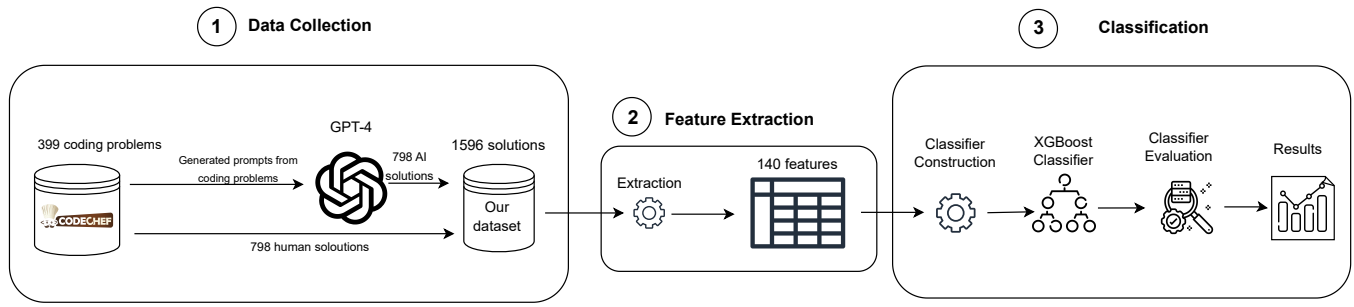


Figure 2: An overview of our approach in detecting GPT-4 generated code

[6, 11, 16, 23, 28, 49] and plagiarism detection [19, 43, 50]. Although these terms are sometimes used interchangeably, authorship attribution deals with identifying the author of code, while plagiarism detection, on the other hand, assumes the author is known and aims to identify instances of unoriginal code [30]. In these studies, different code stylometry features were found to be effective, with layout or typographical features [41] proving to be more accurate than Halstead’s metrics [27],<sup>4</sup> a conventional complexity metric. Some studies used syntactic features [6, 16, 23, 28, 49], whereas others combined layout, lexical, and syntactic features [11, 19, 50].

Our work leverages many of the code stylometry features used in the above studies for a new purpose: to distinguish between human-authored and GPT-4 generated code.

**2.2.3 Machine Learning Approaches.** More generally, machine learning techniques have been applied in various code analysis tasks such as testing [40], defect detection [3], refactoring [34, 48], vulnerability detection [20, 31, 33], program comprehension [47], code smells detection [44], authorship attribution [6, 11, 16, 23, 28, 49], and plagiarism detection [19, 43, 50]. In our work, we use a machine learning technique to determine whether or not a program is GPT-4 generated.

### 3 STUDY DESIGN

In this section, we describe the different phases of our approach, including data collection, feature extraction, and classification. These phases are illustrated in Figure 2.

#### 3.1 Data collection

The data collection phase is depicted on the left of Figure 2. We collected Python problems and human solutions from a repository of programming problems, and we used an AI assistant to generate AI solutions. We chose Python specifically due to its status as a beginner-friendly language [9] commonly adopted in introductory programming courses [35]. To ensure a wide range of programming problems, we chose CodeChef as our problem repository. CodeChef is a renowned competitive coding platform known for offering problems of varying difficulty levels. We believe that there are not inherent differences across other competitive programming platforms, so our approach’s performance on CodeChef should be similar to its performance on these other platforms.

<sup>4</sup>Halstead’s metrics are measurable properties based on the author’s hypothesis that the structure of code is based on two independent properties—operators and operands.

Table 1: Difficulty Levels of Selected CodeChef Problems

Level	Range	Count
Beginner	0 - 999	12
1* Beginner	1000 - 1199	45
1* Advanced	1200 - 1399	71
2* Beginner	1400 - 1499	55
2* Advanced	1500 - 1599	56
3* Beginner	1600 - 1699	60
3* Advanced	1700 - 1799	53
4*	1800 - 1999	30
5*	2000 - 2199	14
6*	2200 - 2499	2
7*	2500 - 5000	1
		399

Due to the absence of public APIs, we extracted data by scraping CodeChef’s website, obtaining both problem sets and user submissions. The data collection process is divided into two steps, as highlighted in Figure 2. We briefly describe each step below.

**Problem Set and Human Solutions Extraction.** In this step, we curated a problem set of coding problems from CodeChef and their corresponding human solutions.

CodeChef assigns each problem on its platform a difficulty score and classifies ranges of difficulties into 11 buckets, as depicted in Table 1. To ensure that our study’s problem set has a good distribution with respect to difficulty, we fetched the 100 most popular problems from each difficulty level. Here popularity refers to the number of accepted solution attempts that existed when the data was scraped (November 2023). In the case that a user submits multiple solutions to the same problem statement, we collect only the latest correct solution submitted by the user. Thus, we began our filtering process with 1100 problem statements.

To further refine our problem set, we selected those problems with at least two correct solutions submitted in 2020. This year was specifically chosen as it postdates the end-of-life for Python 2 in January 2020 and predates the release of AI assistants like Copilot in October 2021 and ChatGPT in November 2022. We purposefully excluded solutions submitted after the release of these AI assistants to ensure that the human-authored solutions in our dataset are not

**Table 2: Final Problem Set Binned into 3 Classes of Difficulty**

Difficulty	Difficulty Scores - Range	Average	Count
Easy	828 - 1417	1224.95	133
Medium	1419 - 1646	1529.51	133
Hard	1647 - 3420	1827.50	133

polluted with AI-generated code. We also purposefully excluded incorrect solutions to avoid including incomplete submissions that do not reflect a typical attempt at a successful solution. In the context of our study, a solution is deemed *correct* if it passes CodeChef’s public tests.<sup>5</sup> Such a solution represents code that is successfully interpreted and executed, effectively solving the specified problem as far as the public tests are concerned.

After this filtering, the number of problems in our dataset was reduced to 419. Upon closer inspection, we found 20 problems that were tagged as “Python3” problems but had only Python2 solutions. These problems were therefore excluded from our study, resulting in a problem set comprising 399 problems, each with at least two correct human-authored Python3 solutions submitted in 2020. In addition, each problem includes comprehensive details from the platform such as the problem statement, unique problem code ID, input and output formats, assigned difficulty score, subtasks, constraints, problem names, user-assigned tags, computed tags, and sample test cases containing input, output, and explanations.

After filtering we noted that we were left with very few problems in the highest levels of difficulty as shown in Table 1. This could be attributed to the fact that competitive programmers often choose C/C++ over Python for various reasons, particularly in contests involving problems of higher difficulty. To mitigate issues with inference due to disparities in the number of problems per category, we re-binned the 399 problems into three classes of difficulty (easy, medium, and hard) of equal size. The classes contain 133 problems each, with average difficulty scores of 1224.95, 1529.51, and 1827.50 respectively as shown in Table 2. This reclassification of difficulty attempts to ensure a balanced representation across categories and enables a fair evaluation of how well our classifiers fare with respect to problems of different difficulty.

**Not including comments.** Although comment-based features like `commentsDensity` [11], `inlineCommentsDensity` [41], and `blockedCommentsDensity` [41] have been used in past works on authorship attribution, we have chosen not to include comments for two reasons:

- (1) Comments may make it too easy for a classifier to determine if the code is human-authored or GPT-4 generated. This is because when you specifically ask for comments from a model like GPT-4, the number of comments is far more than any human would normally write. By excluding comments we handicap our approach and thus provide a lower bound for our classifier. Besides this, comment-based features are easily gameable.

<sup>5</sup>In the context of competitive programming, a submission is deemed *correct* if it passes all public and private test cases; however, we had access only to CodeChef’s public tests for each problem.

You are an expert Python Programmer. Your job is to look at a programming puzzle provided by the user and output 2 different ways to solve the solution in python.

The Input is provided with the following contents:

{The problem statement}

{How the input would be formatted},

{Format to be followed in the output generated},

{Constraints on the variables specified in the problem}

Make sure to take the input from the user considering the input format Output should be printed as defined in the output format

Do not attempt to explain the solution only output the code in the following format:

[PYTHON1]

{Solution to given puzzle in Python}

[\PYTHON1]

[PYTHON2]

{Alternate solution to given puzzle in Python}

[\PYTHON2]

**Figure 3: Prompt used for generating 2 AI code solutions**

- (2) The amount of comments present in GPT-4 generated code varies by the prompt that we give. If we simply query the API of the model to write a program for the problem at hand, the model generates code only and no comments. Alternatively, if we prompt the model to explain the code, then almost every line of code is commented on. Given that the focus of this work is to explore whether we can differentiate between human-authored and GPT-4 generated code, we did not want the choice of prompt to be a variable of the experiment.

Hence we explicitly asked the GPT-4 model not to explain the code and then we stripped any comments that may have been included even by mistake. We apply the same comment-stripping technique to remove any comments from the human code as well. Therefore in this study, we explicitly avoid using comment-related features for classification.

**AI Solutions Generation.** To generate AI solutions to the problems in our problem set we used GPT-4 (Version 0613), which is one of the most powerful and easily accessible generative models available to consumers as of November 2023 through OpenAI [29]. We set the temperature to 0 so that our results are reproducible (at the time of this writing, setting a seed for consistent generation was not available through the API). We used the prompt shown in Figure 3 to obtain two GPT-4 solutions for each of the 399 problems in our problem set, resulting in 798 unique solutions. In constructing our prompt, we employed strategies recommended by OpenAI for effective prompt engineering.<sup>6</sup> The specific strategies we followed are outlined below:

- **Include Details in Your Query to Get More Relevant Answers:** The prompt specifies the details of the task by defining the format of the input (problem statement, input

<sup>6</sup><https://platform.openai.com/docs/guides/prompt-engineering/strategy-provide-reference-text>

format, output format, constraints) and the required output (two Python solutions). This helps in getting relevant and specific answers tailored to the given programming problem.

- **Ask the Model to Adopt a Persona:** The prompt begins with "You are an expert Python Programmer." This tactic of persona adoption sets a context for the responses expected and guides the AI to frame its responses within the expertise of a Python programmer.
- **Use Delimiters to Clearly Indicate Distinct Parts of the Input:** The prompt uses a structured format with clear delimiters, such as [PYTHON1] and [PYTHON2], to separate the two different solutions. This helps the AI understand that two distinct solutions are required and organizes the output in a clear, readable manner. It also enables us to programmatically process the outputs generated.
- **Specify the Steps Required to Complete a Task:** While the prompt implicitly suggests the steps (understand the problem, code the solution), it doesn't explicitly break down the task into smaller steps. In tasks like programming, outlining steps such as analyzing the problem, considering algorithms, and then coding can enhance the quality of the response.

We opted for a zero-shot inference approach. We intentionally did not constrain the output length within the prompt or provide a detailed step-by-step breakdown, among the other suggested strategies to accommodate the diverse nature of problems in our dataset. While this prompt could be further refined, our goal was to develop a pragmatic, 'best effort' prompt reflective of what a typical user might employ.

All generated GPT-4 solutions were syntactically valid and could be successfully parsed, which was important for AST-based features. In the case of duplicate solutions to a problem, we reran the prompt to obtain a new solution to swap in.

Private tests for the problems could not be scraped from the platform, thus we evaluated the GPT-4 generated solutions on the available public tests to check whether they were correct to some degree. We found that only 137 problems had two GPT-4 solutions that satisfy the available public tests, and another 24 problems had only one of the solutions passing the test cases. We used this information to create a sanitized set of 161 problems that includes correct GPT-4 solutions for each problem and an equal number of unique and correct human solutions to those problems for RQ3.

### 3.2 Feature Extraction

Our study leverages a combination of layout, syntactic, and lexical features that have been effectively used in previous studies for authorship attribution and plagiarism detection among human programmers [11, 18, 21, 25]. Layout features refer to the visual organization of code, such as indentation and spacing. Lexical features, on the other hand, are derived from analyzing the tokens within the code, capturing elements such as keywords and literals. While syntactic features are extracted based on the code's structural patterns, involving the arrangement and relationships between various code elements. Our study also incorporates Halstead's metrics [27],

which have been used in previous studies for authorship attribution [8, 41]. We also included additional complexity metrics such as maintainability index [15] and cyclomatic complexity [39] to enrich our approach.

In the feature extraction phase (shown in the middle of Figure 2), we iterate through the Python solution files, systematically generating these code stylometry and complexity features essential for training and evaluating our classifier.

We extracted 31 base features (shown in Table 3) plus variants, leading to 140 features. Most base features have no variants. Feature keywordsDensity has 28 variants, out of 35 Python keywords; these are listed in the replication package. Features ASTNodeTypesTF and ASTNodeTypeAvgDep each have 42 variants, out of 130 possible AST node types; these are listed in the replication package. The prefixes "nttf\_" for ASTNodeTypesTF, "ntad\_" for ASTNodeTypeAvgDep, and the suffix "\_Density" for KeywordsDensity were adopted to correlate variants with their respective base features.

Our approach to feature extraction varies somewhat from prior studies [11, 18, 41]. We normalized our features by source lines of code rather than by character count [11] or by omitting normalization altogether [18, 41]. Additionally, we refrained from logarithmic transformations of some features, as practiced by Caliskan et al. [11], to facilitate the ease of interpretability of our feature set, particularly for visual analysis. For the nestingDepth feature, we considered node types rather than actual tokens. The MaintainabilityIndex feature measures code maintainability by evaluating complexity and modularity and is calculated using complexity metrics such as Cyclomatic Complexity, and SLOC. Our version was computed with the Radon Python library,<sup>7</sup> which uses a modified formula different from the study by Coleman et al. [15]. Although our dataset contains only single-file solutions, the MaintainabilityIndex feature is included in our study as it may yield insights into the relative maintainability of code produced by AI assistants compared to code authored by humans, potentially impacting the performance of our classifier.

Of the 140 features extracted, four are Halstead metrics, selected to explore their viability in this context. Their ineffectiveness in the context of authorship attribution in human code has been pointed out by Berghel and Sallach [8] and Oman and Cook [41], but we included them for completeness. Our research is directed at assessing how well these metrics can identify GPT-4 generated code.

### 3.3 Classification

The righthand side of Figure 2 provides an overview of the classification phase, where we construct a model to distinguish between GPT-4 generated and human-authored code and subsequently evaluate its performance. We describe each step of this phase below.

**Classifier Construction.** To construct our classifier, we chose XGBoost [12], because it is an effective and scalable machine learning algorithm. Additionally, in the study by Bukhari et al [10], XGBoost with syntactic and lexical features had the best performance when considering accuracy and F1 score. This allows us to compare our approach with the best of the earlier study's approaches. XGBoost constructs decision trees iteratively, refining the model by correcting misclassifications at each step. During training, the

<sup>7</sup><https://pypi.org/project/radon/>

**Table 3: Code Stylometry and Code Complexity Features**

Feature	Description
ASTNodeTypesTF [11]	Term frequency of 130 possible AST node types excluding leaves
ASTNodeTypeAvgDep [11]	Average depth of 130 possible AST node types excluding leaves.
avgFunctionLength [21]	The average length of lines in a function.
avgIdentifierLength [21]	The average length of identifier names.
avgLineLength [11]	The average length of characters in each line.
avgParams [11]	The average number of parameters across all functions.
branchingFactor [11]	Average branching factor of the code's AST.
cyclomaticComplexity [39]	The number of decisions within a block of code.
emptyLinesDensity [11]	The number of empty lines divided by source code lines.
keywordsDensity [11]	Frequency of Python keywords divided by source lines of code.
maintainabilityIndex [15]	A metric that gauges the ease of supporting and modifying the source code.
maxDecisionTokens	The maximum number of tokens in decision conditions excluding ternary conditions.
maxDepthASTNode [11]	Maximum depth of an AST node.
nestingDepth [11]	Deepest level to which conditional statements, loops, and functions are nested within each other.
numAssignmentStmtDensity [18]	The total number of assignment statements divided by source code lines.
numClassesDensity	The total number of classes divided by source code lines.
numFunctionCallsDensity [18]	The total number of function calls divided by source code lines.
numFunctionsDensity [18]	The number of functions divided by source code lines.
numInputStmtsDensity [18]	The total number of input statements divided by source code lines.
numKeywordsDensity [11]	The total number of unique Python keywords divided by source code lines.
numLiteralsDensity [11]	The number of literals divided by sloc.
numStatementsDensity [18]	The total number of statements divided by source code lines.
numVariablesDensity [18]	The total number of assignment variables divided by source code lines.
numberOfDistinctOperands [27]	The number of distinct operands.
numberOfDistinctOperators [27]	The number of distinct operators.
sloc [25]	The total number of source code lines.
stdDevLineLength [11]	The standard deviation of character lengths of each line.
stdDevNumParams [11]	The standard deviation of the number of parameters across all functions.
totalNumberOfOperands [27]	The total number of operands.
totalNumberOfOperators [27]	The total number of operators.
whiteSpaceRatio [11]	The ratio of whitespace characters to non-whitespace characters.

algorithm optimizes an objective function to strike a balance between prediction accuracy and model simplicity. At each tree, the algorithm assigns scores to examples, and each example's final prediction is calculated by summing the scores [12]. Through this aggregation process, the resulting model classifies code as either human-authored or GPT-4 generated.

**Classifier Evaluation.** To evaluate the classifier's performance on unseen data, we employ ten-fold cross-validation, which divides the dataset into ten subsets. This approach provides robust assessments by training on nine subsets and testing on the remaining subset. This process is iterated ten times, guaranteeing that each subset serves as the test set exactly once.

To enhance the evaluation process and avoid data poisoning during training, we grouped solutions based on the specific coding problem they addressed, resulting in 399 distinct groups corresponding to the 399 coding problems in our dataset. We employed GroupKFold to ensure that each group, representing solutions to a particular problem, appeared only once in the test set across all folds. This grouping strategy maintains the integrity of the evaluation by preserving the context of solutions within each problem, preventing the model from training and testing on the same problem sets since each problem set contains multiple solutions.

The model's performance is measured in terms of accuracy, recall, precision, F-measure, and AUC-ROC.

**Baselines Comparison.** In our model comparison, we evaluate our approach alongside two baselines: (1) a naive baseline approach, based on the assumption that GPT-4 generated code can be detected through random guessing, and (2) the approach presented by Bukhari et al. [10] that identified AI-generated solutions for C programming assignments. To benchmark our classifier, we replicated the methodology of Bukhari et al. [10] using Python and an XGBoost classifier, motivated by XGBoost performing the best in their study on syntactic and lexical features. We evaluate the performance of our classifier using metrics such as accuracy, recall, precision, F-measure, and AUC-ROC.

Given that our case study focuses solely on utilizing the CodeChef dataset, we will not employ statistical significance tests like the Mann-Whitney U or T-test between our approach and the second baseline. These tests typically require multiple datasets, a requirement our single-dataset study does not meet. Additionally, such tests cannot be employed on the classifier's raw prediction scores due to the presence of high probability misclassifications. Furthermore, applying statistical significance tests to performance metrics derived from ten-fold cross-validation is problematic due to the overlap of the training dataset across folds, with about 80% of data shared between each pair of training sets [17]. This overlap leads to interdependence among the folds, thereby violating the independence assumption of these tests. As a result, we risk incorrectly

**Table 4: Classifier Performance Comparison among Different Approaches for Distinguishing between AI-generated and Human-authored Code**

	Our Approach		Baseline		
	All	Non-Gameable	Naive	n-grams + L n = 2	n = 3
Accuracy	0.91	0.89	-	0.86	0.88
Precision	0.91	0.89	0.5	0.86	0.87
Recall	0.91	0.89	0.5	0.88	0.88
F1-score	0.91	0.89	0.5	0.87	0.88
AUC-ROC	0.91	0.89	-	0.86	0.88

rejecting the null hypothesis even in instances where there may be no actual difference.

## 4 RESULTS

We present the results of our model evaluation in relation to our research questions. For each research question, we first outline our approach to addressing it, followed by the observed results.

**(RQ1.)** *How well can code-stylometry features distinguish human-authored code from GPT-4 generated code?*

**Approach.** To address our research question, we build a classifier trained on code stylometry features. After training, we evaluate its performance using ten-fold cross-validation, focusing on metrics such as precision, recall, F-measure, and AUC-ROC. High scores in these metrics will support our research hypothesis, demonstrating our approach’s ability to distinguish human-authored code from AI-generated code. Additionally, we compare this classifier with an alternative classifier trained on the same feature set but augmented with Halstead’s metrics. The objective is to evaluate the impact of Halstead’s metrics in detecting GPT-4 generated code. This will involve a comparative analysis of the classifiers’ performances with and without Halstead’s metrics. Moreover, we contrast the performance of our classifier with two baselines.

The Naive Baseline is a random guess and its performance metrics can be computed by applying statistics to our dataset. The precision is calculated by dividing the number of GPT-4 generated code by the total number of solutions:

$$\text{Precision} = \frac{\text{number of GPT-4 generated code}}{\text{total number of solutions}} = 0.5 \quad (1)$$

The recall is 0.5, reflecting the classifier’s two possible outcomes—identifying code as either GPT-4 generated or human-authored. This results in a probability of 0.5 for classifying solutions as GPT-4 generated. Based on precision and recall values, we compute the  $F_1$  score of the naive baseline as:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 0.5 \quad (2)$$

**Results.** As shown in Table 4,<sup>8</sup> our approach achieved a high average precision and recall of 0.91, ensuring accurate and comprehensive identification of GPT-4 generated code. This highlights the potential of code stylometry in differentiating between GPT-4 generated and human-authored code. Comparing the classifiers, one with and the other without Halstead metrics, we found a striking similarity in their performance metrics. Between both classifiers, all the metrics considered were the same, except recall which was higher for the classifier with Halstead metrics by 0.01. This observation suggests that the presence of Halstead metrics does not considerably enhance the classifier’s ability to distinguish between human-authored and GPT-4 generated code supporting past work [25, 37].

When compared to the baselines, our classifier shows a considerable improvement. It notably outperformed the Naive Baseline, which has a precision and recall of 0.5, demonstrating that our classifier considerably exceeds what would essentially be random guessing. We also compare with the work presented in Bukhari et al. [10] that incorporated lexical features of 2-4 n-grams. However, in our replication, we could only process 2-3 n-grams due to the memory-intensive nature of the task. The data for the 4-gram model was at least 212.33 GBs, resulting in out-of-memory errors in our machine (Macbook Air 2020, with 16GBs of RAM). Also, it took more than 8 hours to extract the data for the 4-gram model. However, our classifier has no such issue. We were also able to achieve higher precision and recall by 4% and 3% respectively.

Our study demonstrates a higher predictive power over the n-grams baseline and provides interpretable predictions. Unlike its n-gram-based features, which may obscure the reasoning behind predictions, our model uses a feature set that clarifies the decision-making process. To understand the influence of specific features on our model’s predictions, we used the SHAP framework [38], a method renowned for its interpretability of machine learning models. SHAP offers tools for both local (individual) and global (overall) explanations of model predictions.

The global interpretive power of our classifier is demonstrated in the SHAP summary plot depicted in Figure 4. This plot visualizes key features in our classifier, arranging them on the y-axis by their aggregate SHAP values, with the highest at the top. The x-axis displays these SHAP values, showing how each feature shifts the prediction from a neutral base value, indicated by the vertical line at 0 on the x-axis. Deviations to the left or right increase the likelihood of the prediction being a human or GPT-4 class, respectively. The plot uses a blue-to-red color gradient to signify feature magnitudes, and the data points represent feature values across instances.

As depicted in Figure 4, the *avgLineLength* is the most important feature, distinctly separating human and GPT-4 classes where its lower values are typically associated with the human class, whereas higher values are associated with the GPT-4 class. This implies that a line of code from a human is shorter on average compared to a line of code from GPT-4. In contrast, the *ntad\_Name* feature, is the tenth most important feature and quantifies the average depth at which the *Name* node occurs within an AST. The *Name* node in

<sup>8</sup>It is just a coincidence that the performance metrics for our models all have the same value when considering all features (0.91) and all non-gameable features (0.89)

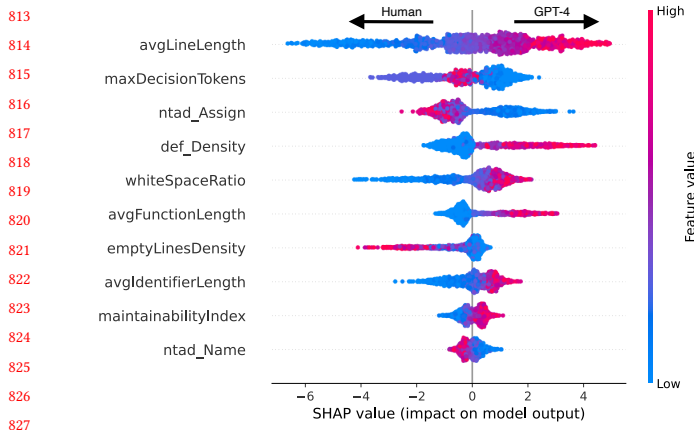


Figure 4: SHAP feature importance of our approach



Figure 5: SHAP feature importance of non-gameable features

Python’s AST represents identifiers, which are the names of variables, functions, classes, modules, or other objects in the code. The distribution of the *ntad\_Name* feature across all instances shows a relatively narrow range of SHAP values, suggesting that its impact on the model is less compared to the features ranked above it. It is important to note that SHAP functions as an explanation model that provides an interpretable approximation of our classifier. Therefore, while SHAP values offer a simplified and interpretable view, the feature importance rankings derived directly from our classifier are based on the classifier’s internal mechanisms. Consequently, the feature importance rankings from our classifier, do not entirely align with SHAP’s. Despite these differences, there is considerable overlap in the primary features identified by both methods. This overlap highlights the value of SHAP in interpreting the model’s predictions, offering insights into how features influence outcomes rather than detailing the classifier’s internal mechanisms.

**(RQ2.)** How influential are non-gameable features in differentiating human-authored vs. GPT-4 generated code?

**Approach.** To assess the impact of non-gameable features on the detection of GPT-4 generated code, we build a classifier that excludes gameable features.

In our analysis, we consider the non-code layout features *emptyLinesDensity* and *whiteSpaceRatio* as gameable features. After training on the non-gameable features, we evaluate the performance of this classifier in contrast to our classifier with both gameable and non-gameable features. We also evaluate this classifier with the same n-grams baseline compared in RQ1 as the baseline does not include the features we consider gameable in its feature set. Through these comparisons, we aim to evaluate the relative importance of using only non-gameable features in the classification of code as human-authored or GPT-4 generated.

**Results.** As shown in Table 4, there is a noticeable but not severe drop in performance for the non-gameable classifier compared to the classifier built on gameable and non-gameable features. This suggests that although gameable features contribute to the classifier’s accuracy, non-gameable features alone still provide a high

Table 5: Classifier Performance Comparison on Correct and Randomly Sampled Solutions

	Our Approach		Baseline (n-grams + L)			
	C	R	n = 2		n = 3	
			C	R	C	R
Accuracy	0.86	0.87	0.83	0.84	0.87	0.86
Precision	0.87	0.87	0.83	0.84	0.87	0.86
Recall	0.86	0.88	0.81	0.85	0.87	0.85
F1-score	0.86	0.87	0.82	0.84	0.87	0.86
AUC-ROC	0.86	0.87	0.83	0.84	0.87	0.86

C = Correct Solutions, R = Random Solutions, L = Lexical Features

predictive power. In comparison to the n-grams baseline, the non-gameable classifier still performs better and is interpretable as evident in the summary plot of Figure 5. The plot reveals that aside from the two gameable features among the ten most important features of the classifier trained on both gameable and non-gameable features shown in Figure 4, the other features remain consistent with a slight reordering. The absence of these two gameable features accounts for the performance dip in the non-gameable classifier. Consequently, *stdDevLineLength* and *nttf\_Name*, the latter representing the term frequency of the Name node, now appear in the top ten. The *stdDevLineLength* feature influences predictions towards the GPT-4 class at lower values and towards the human class at intermediate values. Conversely, *nttf\_Name* influences predictions towards the human class at higher values and towards the GPT-4 class when lower.

**(RQ3.)** How well does the classifier perform when trained and evaluated on only correct solutions?

**Approach.** To address this research question, we refine our dataset to include only correct solutions, resulting in a balanced dataset of 596 correct solutions from both humans and GPT-4. We aim to evaluate our classifier’s predictive power by eliminating the potential noise incorrect solutions might introduce. This ensures that the detection of GPT-4 generated solutions is based on inherent characteristics of code, not the errors they might contain.



**Results.** Table 5 shows a slight decline in performance when compared to the classifier in Table 4. The minor decline could stem from the correct solutions being a smaller subset. Hence, we randomly sample the solutions comprising both correct and incorrect solutions with the same distribution of difficulty levels. When compared with the random solutions classifier, both perform almost the same. This implies that genuine differences in coding styles between human and GPT-4 generated code are being detected, rather than errors introduced by incorrect solutions. When compared to the baseline [10], for correct solutions, our classifier outperformed the 2-gram baseline but was marginally less effective than the 3-gram. The classifier with randomly sampled solutions showed noticeable improvement over the 2-gram baseline and was marginally better than the 3-gram.

**(RQ4.)** *How well does the classifier perform when trained and evaluated across varying levels of problem difficulty?*

**Approach.** To address this research question, we construct separate classifiers for the three problem difficulty levels outlined in Table 2. By training and evaluating these classifiers independently, we evaluate their performance and the potential impact of problem complexity on the classifier’s ability to correctly identify GPT-4 generated code. This stratified analysis allows us to understand the nuances of classifier performance across problem difficulty levels.

**Results.** Table 6 shows a correlation between the classifier performance and the difficulty of the problems. The performance of classifiers of easy and medium problem difficulty are close as they have the same F1-score of 0.89. The performance of the classifier with hard problems had a minor drop in performance with an F1-score of 0.87. When compared to the baseline, there is no considerable difference in the baseline for easy questions. For both medium and hard questions, our classifiers perform better than both 2-gram and 3-gram classifiers, showing improvements of 3% and 9%, respectively. This result highlights the effectiveness of our approach.

## 5 DISCUSSION

Based on our findings, we make observations on our classifier, examining why it correctly and incorrectly predicts solutions in specific instances. Understanding the reason behind its performance on a specific instance is crucial for humans to make the final decision whether it is human-authored or GPT-4 generated.

### 5.1 Correctly Predicted Solution

Figures 6 and 7 present examples where our model correctly classifies (i.e., correctly distinguishes between human-authored and GPT-4 generated code). Examining the important features in Figure 4 and 5, we can gain insights into the model’s overall decision-making process by investigating the waterfall plot for each example. The waterfall plot presents the model’s expected value, with each row representing how each feature contributes positively (red) towards GPT-4 generated code or negatively (blue) towards human-authored code.

Figure 6, shows a correctly predicted human-authored code and its corresponding SHAP waterfall plot. This plot provides a local explanation of the classifier, visually depicting the key features that influence individual predictions, starting from the base value.

At the bottom of the plot in Figure 6b, the model’s prediction begins at this base value. It progresses towards the GPT-4 generated class, influenced by the cumulative effect of 127 other features whose contributions are relatively minor and thus aggregated in the plot. These are ranked in ascending order of SHAP values. The *ntad\_Assign* feature, representing the AST node for assignment operations, has the most impact towards the GPT-4 class, with its value of 1.29. The value of this feature, in addition to the value of other features with red arrows in the plot, is indicative of the GPT-4 generated class. However, the most impactful feature overall is the *avgLineLength*, its value of 14.36 in addition to the value of other features with blue arrows in the plot influences the model’s decision towards its final prediction, the human class. This shows that code having an *avgLineLength* (the average length of characters in each line) of 14.36, *emptyLinesDensity*, (i.e. the ratio of empty lines to sloc) of 0.4, *def\_Density* of 0, *whiteSpaceRatio* (i.e. the ratio of whitespace to non-whitespace characters) of 0.35, *avgIdentifierLength* (i.e. the average length of identifier names) of 2.67 and no function is likely human-authored, not taking into consideration the other features.

Figure 7 shows a correctly predicted GPT-4 generated code and its corresponding SHAP waterfall plot. In Figure 7b, the most important feature with a value of 0.25 is the *def\_Density*. The value of this feature alongside the value of other features with red arrows influences the model’s decision towards the final prediction, the GPT-4 class. This suggests that code having *def\_Density* of 0.25, *ntad\_Assign* of 1, *maxDecisionTokens* of 0, *maintainabilityIndex* of 81.856 is likely GPT-4 generated, not taking into consideration other features. *MaintainabilityIndex* is a feature that isn’t observed by looking at the code, this is an interesting find on how a complexity metric could potentially influence the classifier’s decision.

These findings underscore the model’s ability to correctly predict solutions based on distinctive features, shedding light on the significance of specific code characteristics in the classification process.

### 5.2 Incorrectly Predicted Solution

Figure 8 presents an incorrectly predicted GPT-4 generated code and its corresponding SHAP waterfall plot. This case is concerning, as the model incorrectly guesses the authorship of the code as GPT-4 generated. While a false negative (predicting GPT-4 generated code as human-authored) is bad, we especially want to avoid false positives. This is because we do not want to unjustly accuse someone of presenting GPT-4 generated code when they have shared their written code. However, all classification techniques are bound to have some false positives. We hope that our choice of an explainable model helps educators look at the SHAP waterfall plot and understand the reasons before making a final decision as an educator.

By examining important features from SHAP in Figure 4 and 5, we aim to gain insights into why the model made this incorrect prediction. In Figure 8b, the two top features that influence the model’s prediction towards the GPT-4 generated class are related to the length of the code (i.e., *avgFunctionLength*, *avgLineLength*) and have values of 13 and 24.86, respectively. In the case of both *avgLineLength* and *avgFunctionLength*, their values are high compared

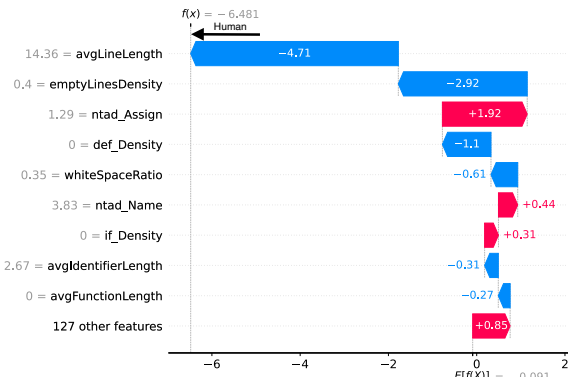
**Table 6: Classifier Performance Comparison Across Levels of Problem Difficulty**

	Our Approach			Baseline (n-grams + Lexical Features)					
	Easy	Medium	Hard	n = 2			n = 3		
				Easy	Medium	Hard	Easy	Medium	Hard
Accuracy	0.89	0.89	0.87	0.87	0.79	0.80	0.89	0.86	0.80
Precision	0.87	0.88	0.89	0.85	0.80	0.79	0.89	0.87	0.80
Recall	0.91	0.90	0.86	0.89	0.77	0.82	0.88	0.85	0.81
F1-score	0.89	0.89	0.87	0.87	0.79	0.80	0.89	0.86	0.80
AUC-ROC	0.89	0.89	0.87	0.87	0.79	0.80	0.89	0.86	0.80

```

1 n, m = map(int, input().split())
2 mi = 2
3 ma = n + m
4 ans = [1 for i in range(ma + 1)]
5
6 for i in range(2, int(ma**0.5) + 1):
7     for j in range(i + i, ma + 1, i):
8         ans[j] = 0
9 ans[0] = 0
10 ans[1] = 0
11 print(ans.count(1))
12
13
14
    
```

(a) Human code



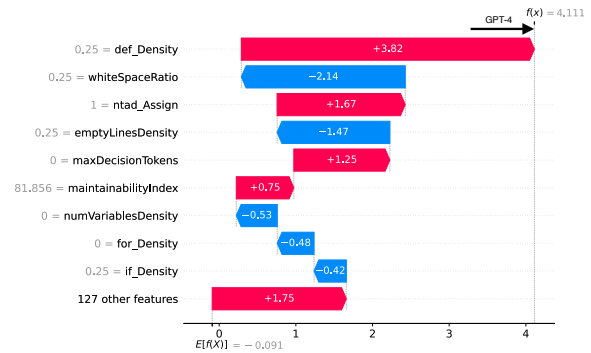
(b) SHAP waterfall plot showing how features impact the model's decision for this code

**Figure 6: Correctly predicted human code**

```

1 def solve(n, m):
2     return 1 if min(n, m) > 1 else 2
3
4 n, m = map(int, input().strip().split())
5 print(solve(n, m))
    
```

(a) GPT-4 generated code



(b) SHAP waterfall plot showing how features impact the model's decision for this code

**Figure 7: Correctly predicted GPT-4 generated code.**

to the feature values of other observations within the dataset, and features with higher *avgLineLength* and *avgFunctionLength* tend to drive the model's prediction towards the GPT-4 generated class.

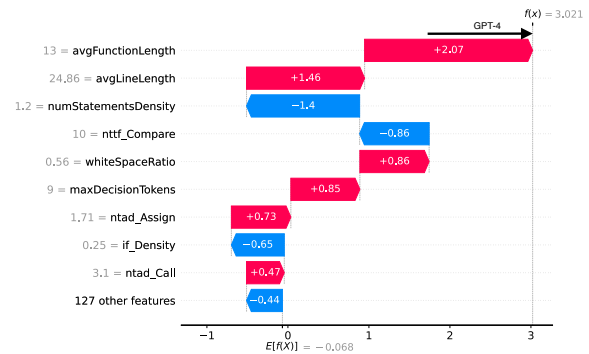
These observations explain why the model predicted this code as GPT-4 generated, highlighting the challenges in accurately distinguishing certain code characteristics and the potential consequences of false positives.

```

1161 1 def getCount(h,m,i) :
1162 2     h=int(h)
1163 3     m=int(m)
1164 4     h1=0
1165 5     lst=[11,22,33,44,55,66,77,88,99]
1166 6     while(h1<h) :
1167 7         for m1 in range(0,m) :
1168 8             if h1<10 :
1169 9                 if(m1<10 and h1==m1) : count[i]+=1
1170 10                if (m1 in lst and m1%10==h1) : count[i]+=1
1171 11            else :
1172 12                if(m1 in lst and h1==m1) : count[i]+=1
1173 13                if(h1 in lst and h1%10==m1) : count[i]+=1
1174 14            h1+=1
1175 15        t=int(input())
1176 16        count=[0]*t
1177 17        for i in range(0,t) :
1178 18            h,m=input().split()
1179 19            getCount(h,m,i)
1180 20
1181 21        for i in count : print(i)

```

(a) Human code



(b) SHAP waterfall plot showing how features impact the model's decision for this code

Figure 8: Human code predicted incorrectly as AI-generated code

## 6 THREAT TO VALIDITY

We break down the threats into two parts, external and construct.

### 6.1 External Threat to Validity

These threats relate to the ability to generalize based on our results. In this study, we conducted an empirical investigation on the competitive programming platform CodeChef for human-authored code and utilized GPT-4 for AI-generated code. Coding practices may vary on other platforms. However, in this study, to encompass different styles, we selected the most popular 100 problems for each difficulty level (i.e., from beginner level to 7\* level) and extracted those that include Python solutions. This approach should enable us to generalize coding styles from beginners to experts. Our classifier performed well on easy, medium, and hard questions and we believe that due to the classifier's performance on easy and medium questions, it can perform fairly well on introductory programming courses. This is because problems in programming courses are likely to be of easy to medium difficulty. However, human solutions provided for easy to medium questions may not accurately represent the expertise level of all users. This means the solutions, while insightful, might not fully reflect the skill level of specific groups, such as students new to programming in introductory programming courses. Nonetheless, they do offer a general insight into human coding styles. Another concern is since we only generate code using GPT-4, the generated code may not be representative of code by other AI assistants. However, compared to other tools, GPT-4 is the most popular AI assistant and should represent real-world usage. In future work, we aim to expand to other programming languages, coding platforms and adapt to a broader range of AI models.

### 6.2 Construct Validity

These threats relate to the degree to which our measurements are captured. In terms of correctness of AI-generated code, we extract public test cases from CodeChef, which do not include private tests for the problems. We decided not to pursue direct submissions of AI-generated code to CodeChef, as this would violate ethical guidelines to submit AI-generated code as a human solution.

## 7 CONCLUSION

The advent of AI assistants has introduced a new form of academic dishonesty, where students submit AI-generated code as their work. In this study, we investigated the impact of using code stylometry features to differentiate between human-authored and GPT-4 generated code, focusing on submissions from CodeChef and GPT-4 generated solutions in Python. The findings demonstrate our approach's promise. Our classifier achieved an F1-score and AUC-ROC score of 0.91, highlighting its potential as a preliminary tool for identifying AI-generated code. Additionally, we identified several key distinguishing features, with the average line length as the most important feature. By providing a means to identify GPT-4 generated code, our study contributes to the ongoing discourse on the use and regulation of AI assistance in coding tasks.

**Future Work.** To broaden the scope and applicability of our study, we could evaluate our hypothesis that our approach can perform as well on Python programs from other sources such as other competitive programming platforms besides CodeChef and programming assignments from introductory courses. We could also extend our study to include other programming languages. Additionally, further research could evaluate the effectiveness of our approach in identifying AI-generated code that has been intentionally modified post-generation or through prompt engineering. Such investigation would provide valuable insights into the robustness and shortcomings of our detection methods when AI-generated code is deliberately altered to evade identification.

## REFERENCES

- [1] [n. d.]. <https://codequiry.com/>
- [2] [n. d.]. <https://copyleaks.com/>
- [3] Simran Aggarwal. 2020. Software code analysis using ensemble learning techniques. In *Proceedings of the 1st International Conference on Advanced Information Science and System (AISS '19)*. Association for Computing Machinery, New York, NY, USA, 1–7. <https://doi.org/10.1145/3373477.3373486>
- [4] Alex Aiken. [n. d.]. MOSS. <https://theory.stanford.edu/~aiken/moss/>
- [5] Ibrahim Alblawi. [n. d.]. Plagiarism in Programming Assessments: A Systematic Review. *ACM Transactions on Computing Education* 20, 1 ([n. d.]).
- [6] Bander Alsulami, Edwin Dauber, Richard Harang, Spiros Mancoridis, and Rachel Greenstadt. 2017. Source Code Authorship Attribution Using Long Short-Term Memory Based Networks. In *Computer Security – ESORICS 2017*, Simon N. Foley, Dieter Gollmann, and Einar Snekkenes (Eds.). Vol. 10492. Springer International Publishing, Cham, 65–82. [https://doi.org/10.1007/978-3-319-66402-6\\_6](https://doi.org/10.1007/978-3-319-66402-6_6) Series Title: Lecture Notes in Computer Science.
- [7] Brett A. Becker, Paul Denny, James Finnie-Ansley, Andrew Luxton-Reilly, James Prather, and Eddie Antonio Santos. 2023. Programming Is Hard - Or at Least It Used to Be: Educational Opportunities and Challenges of AI Code Generation. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1 (Toronto ON, Canada) (SIGCSE 2023)*. Association for Computing Machinery, New York, NY, USA, 500–506. <https://doi.org/10.1145/3545945.3569759>
- [8] H. L. Berghel and D. L. Sallach. 1984. Measurements of Program Similarity in Identical Task Environments. *SIGPLAN Not.* 19, 8 (aug 1984), 65–76. <https://doi.org/10.1145/988241.988245>
- [9] Jason R Briggs. 2012. *Python for kids: A playful introduction to programming*. no starch press.
- [10] Sufiyan Bukhari, Benjamin Tan, and Lorenzo De Carli. 2023. Distinguishing AI-and Human-Generated Code: a Case Study. (2023).
- [11] Aylin Caliskan-Islam, Richard Harang, Andrew Liu, Arvind Narayanan, Clare Voss, Fabian Yamaguchi, and Rachel Greenstadt. 2015. De-anonymizing Programmers via Code Stylometry. (2015), 255–270.
- [12] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA) (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. <https://doi.org/10.1145/2939672.2939785>
- [13] Robert Clarke and Thomas Lancaster. 2013. Commercial Aspects of Contract Cheating. In *Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education (ITI'13)*. 219–224.
- [14] Coderbyte. 2021. Detect candidates that cheat with AI / ChatGPT. <https://help.coderbyte.com/knowledge/detect-candidates-that-cheat-with-ai-/chatgpt>
- [15] D. Coleman, D. Ash, B. Lowther, and P. Oman. 1994. Using metrics to evaluate software system maintainability. *Computer* 27, 8 (Aug 1994), 44–49. <https://doi.org/10.1109/2.303623>
- [16] Edwin Dauber, Aylin Caliskan, Richard Harang, and Rachel Greenstadt. 2018. Git Blame Who? Stylistic Authorship Attribution of Small, Incomplete Source Code Fragments. In *Proceedings of the 40th International Conference on Software Engineering: Companion Proceedings (Gothenburg, Sweden) (ICSE '18)*. Association for Computing Machinery, New York, NY, USA, 356–357. <https://doi.org/10.1145/3183440.3195007>
- [17] Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10, 7 (1998), 1895–1923.
- [18] John L. Donaldson, Ann-Marie Lancaster, and Paula H. Sposato. 1981. A plagiarism detection system. In *Proceedings of the twelfth SIGCSE technical symposium on Computer science education - SIGCSE '81*. ACM Press, St. Louis, Missouri, United States, 21–25. <https://doi.org/10.1145/800037.800955>
- [19] Wenyuan Dong, Zhiyong Feng, Hua Wei, and Hong Luo. 2020. A Novel Code Stylometry-based Code Clone Detection Strategy. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*. 1516–1521. <https://doi.org/10.1109/IWCMC48107.2020.9148302> ISSN: 2376-6506.
- [20] Mojtaba Eshghie, Cyrille Artho, and Dilian Gurov. 2021. Dynamic Vulnerability Detection on Smart Contracts Using Machine Learning. In *Proceedings of the 25th International Conference on Evaluation and Assessment in Software Engineering (EASE '21)*. Association for Computing Machinery, New York, NY, USA, 305–312. <https://doi.org/10.1145/3463274.3463348>
- [21] J.A.W. Faidhi and S.K. Robinson. 1987. An empirical approach for detecting program similarity and plagiarism within a university programming environment. *Computers Education* 11, 1 (1987), 11–19. [https://doi.org/10.1016/0360-1315\(87\)90042-X](https://doi.org/10.1016/0360-1315(87)90042-X)
- [22] James Finnie-Ansley, Paul Denny, Brett Becker, Andrew Luxton-Reilly, and James Prather. 2022. The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming. 10–19. <https://doi.org/10.1145/3511861.3511863>
- [23] Sophia F. Frankel and Krishnendu Ghosh. 2021. Machine Learning Approaches for Authorship Attribution using Source Code Stylometry. In *2021 IEEE International Conference on Big Data (Big Data)*. 3298–3304. <https://doi.org/10.1109/BigData52589.2021.9671332>
- [24] Github. 2021. *Copilot: Your AI Pair Programmer*. Retrieved 9-October-2023 from <https://github.com/features/copilot>
- [25] Sam Grier. 1981. A tool that detects plagiarism in Pascal programs. *ACM SIGCSE Bulletin* 13, 1 (Feb. 1981), 15–20. <https://doi.org/10.1145/953049.800954> Number: 1.
- [26] HackerRank. 2021. HackerRank Launches AI-Powered Plagiarism Detection. <https://www.hackerrank.com/blog/hackerrank-launches-ai-powered-plagiarism-detection/>
- [27] M. H. Halstead. 1972. Natural laws controlling algorithm structure? *ACM SIGPLAN Notices* 7, 2 (Feb. 1972), 19–26. <https://doi.org/10.1145/953363.953366> Number: 2.
- [28] Pengnan Hao, Zhen Li, Cui Liu, Yu Wen, and Fanming Liu. 2022. Towards Improving Multiple Authorship Attribution of Source Code. In *2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*. 516–526. <https://doi.org/10.1109/QRS57517.2022.00059> ISSN: 2693-9177.
- [29] Cheng Jiao, Neel R Edupuganti, Parth A Patel, Tommy Bui, Veeral Sheth, and Neel Edupuganti. 2023. Evaluating the Artificial Intelligence Performance Growth in Ophthalmic Knowledge. *Cureus* 15, 9 (2023).
- [30] Vaibhavi Kalgutkar, Ratinder Kaur, Hugo Gonzalez, Natalia Stakhanova, and Alina Matyukhina. 2019. Code Authorship Attribution: Methods and Challenges. *Comput. Surveys* 52, 1 (Feb. 2019), 3:1–3:36. <https://doi.org/10.1145/3292577>
- [31] Gurpreet Kaur, Yasir Malik, Hamman Samuel, and Fehmi Jaafar. 2018. Detecting Blind Cross-Site Scripting Attacks Using Machine Learning. In *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*. ACM, Shanghai China, 22–25. <https://doi.org/10.1145/3297067.3297096>
- [32] Majeed Kazemitabaar, Justin Chow, Carl Ka To Ma, Barbara J. Ericson, David Weintrop, and Tovi Grossman. 2023. Studying the Effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'23)*. Article 455, 23 pages.
- [33] Jorrit Kronjee, Arjen Hommersom, and Harald Vranken. 2018. Discovering software vulnerabilities using data-flow analysis and machine learning. In *Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/3230833.3230856>
- [34] Lov Kumar, Shashank Mouli Satapathy, and Lalita Bhanu Murthy. 2019. Method Level Refactoring Prediction on Five Open Source Java Projects using Machine Learning Techniques. In *Proceedings of the 12th Innovations on Software Engineering Conference (formerly known as India Software Engineering Conference) (ISEC'19)*. Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/3299771.3299777>
- [35] Wanda M. Kunkle and Robert B. Allen. 2016. The Impact of Different Teaching Approaches and Languages on Student Learning of Introductory Programming Concepts. *ACM Trans. Comput. Educ.* 16, 1, Article 3 (jan 2016), 26 pages. <https://doi.org/10.1145/2785807>
- [36] Sam Lau and Philip Guo. 2023. From "Ban It Till We Understand It" to "Resistance is Futile": How University Programming Instructors Plan to Adapt as More Students Use AI Code Generation and Explanation Tools Such as ChatGPT and GitHub Copilot. In *Proceedings of the ACM Conference on International Computing Education Research (ICER'23) - Volume 1*. 106–121.
- [37] Ronald J. Leach. 1995. Using Metrics to Evaluate Student Programs. *SIGCSE Bull.* 27, 2 (jun 1995), 41–43. <https://doi.org/10.1145/201998.202010>
- [38] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR abs/1705.07874* (2017). arXiv:1705.07874 <http://arxiv.org/abs/1705.07874>
- [39] T.J. McCabe. 1976. A Complexity Measure. *IEEE Transactions on Software Engineering* SE-2, 4 (Dec. 1976), 308–320. <https://doi.org/10.1109/TSE.1976.233837> Number: 4 Conference Name: IEEE Transactions on Software Engineering.
- [40] Aravind Nair, Karl Meinke, and Sigrid Eldh. 2019. Leveraging mutants for automatic prediction of metamorphic relations using machine learning. In *Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE 2019)*. Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3340482.3342741>
- [41] P. W. Oman and C. R. Cook. 1989. Programming style authorship analysis. In *Proceedings of the seventeenth annual ACM conference on Computer science : Computing trends in the 1990's Computing trends in the 1990's - CSC '89*. ACM Press, Louisville, Kentucky, United States, 320–326. <https://doi.org/10.1145/75427.75469>
- [42] OpenAI. 2022. *Introducing ChatGPT*. Retrieved 9-October-2023 from <https://openai.com/blog/chatgpt>
- [43] Julia Oppen-Rhein, Bastian Küppers, and Ulrik Schroeder. 2019. Requirements for Author Verification in Electronic Computer Science Exams. In *Proceedings of the 11th International Conference on Computer Supported Education*. SCITEPRESS - Science and Technology Publications, Heraklion, Crete, Greece, 432–439. <https://doi.org/10.5220/0007736104320439>

<p>1393 [44] Manjula Peiris and James H. Hill. 2014. Towards detecting software performance          1394 anti-patterns using classification techniques. <i>ACM SIGSOFT Software Engineering</i>          1395 <i>Notes</i> 39, 1 (Feb. 2014), 1–4. <a href="https://doi.org/10.1145/2557833.2560586">https://doi.org/10.1145/2557833.2560586</a>          1396 [45] Ben Puryear and Gina Sprint. 2022. Github copilot in the classroom: learning to          1397 code with AI assistance. <i>Journal of Computing Sciences in Colleges</i> 38, 1 (2022),          1398 37–47.          1399 [46] Amazon Web Services. 2023. <i>What is CodeWhisperer?</i> Retrieved 9-October-          1400 2023 from <a href="https://docs.aws.amazon.com/codewhisperer/latest/userguide/what-is-cwspr.html">https://docs.aws.amazon.com/codewhisperer/latest/userguide/what-</a>          1401 <a href="https://docs.aws.amazon.com/codewhisperer/latest/userguide/what-is-cwspr.html">is-cwspr.html</a>          1402 [47] Zhiyu Sun, Fang Peng, Junrui Guan, and Yanchun Sun. 2019. An Approach to          1403 Helping Developers Learn Open Source Projects Based on Machine Learning. In          1404 <i>Proceedings of the 11th Asia-Pacific Symposium on Internetware</i> (Fukuoka, Japan)          1405 (<i>Internetware '19</i>). Association for Computing Machinery, New York, NY, USA,          1406 Article 13, 10 pages. <a href="https://doi.org/10.1145/3361242.3361251">https://doi.org/10.1145/3361242.3361251</a>          1407          1408          1409          1410          1411          1412          1413          1414          1415          1416          1417          1418          1419          1420          1421          1422          1423          1424          1425          1426          1427          1428          1429          1430          1431          1432          1433          1434          1435          1436          1437          1438          1439          1440          1441          1442          1443          1444          1445          1446          1447          1448          1449          1450</p>	<p>[48] Irene Tollin, Francesca Arcelli Fontana, Marco Zanoni, and Riccardo Roveda. 2017.          Change Prediction through Coding Rules Violations. In <i>Proceedings of the 21st</i>  <i>International Conference on Evaluation and Assessment in Software Engineering</i>  <i>(EASE '17)</i>. Association for Computing Machinery, New York, NY, USA, 61–64.  <a href="https://doi.org/10.1145/3084226.3084282">https://doi.org/10.1145/3084226.3084282</a>          [49] Farhan Ullah, Sohail Jabbar, and Fadi Al-Turjman. 2020. Programmers' de-          anonymization using a hybrid approach of abstract syntax tree and deep learning.  <i>Technological Forecasting and Social Change</i> 159 (Oct. 2020), 120186. <a href="https://doi.org/10.1016/j.techfore.2020.120186">https://doi.org/10.1016/j.techfore.2020.120186</a>          [50] Nickolay Viuginov, Petr Grachev, and Andrey Filchenkov. 2021. A Machine          Learning Based Plagiarism Detection in Source Code. In <i>Proceedings of the 2020</i>  <i>3rd International Conference on Algorithms, Computing and Artificial Intelligence</i>  <i>(ACAI '20)</i>. Association for Computing Machinery, New York, NY, USA, 1–6.  <a href="https://doi.org/10.1145/3446132.3446420">https://doi.org/10.1145/3446132.3446420</a></p>	<p>1451          1452          1453          1454          1455          1456          1457          1458          1459          1460          1461          1462          1463          1464          1465          1466          1467          1468          1469          1470          1471          1472          1473          1474          1475          1476          1477          1478          1479          1480          1481          1482          1483          1484          1485          1486          1487          1488          1489          1490          1491          1492          1493          1494          1495          1496          1497          1498          1499          1500          1501          1502          1503          1504          1505          1506          1507          1508</p>
---	--	---