Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Original Article

Check for updates

A Comparative Analysis of ChatGPT and Traditional Machine Learning Algorithms on Real-World Data

Bnar Kamaran Arif [a, b *], **Aso M. Aladdin** [a, c]

[a] Computer Science Department, College of Science, Charmo University, Chamchamal, Iraq.
[b] Information Technology Department, College of Informatics, Sulaimani Polytechnic University, Sulaymaniyah, Iraq.
[c] Information Technology Department, Tishk International University, Sulaymaniyah, Iraq.

Submitted: 17 May 2025
Revised: 10 June 2025
Accepted: 18 August 2025

* **Corresponding Author**: bnar.kamaran@chu.edu.iq

Keywords: ChatGPT, Algorithm, Machine learning, Accuracy, Time processing.

How to cite this paper: B. K. Arif, A. M. Aladdin, " A Comparative Analysis of ChatGPT and Traditional Machine Learning Algorithms on Real-World Data", KJAR, vol. 10, no. 2, pp: 93-118, Dec 2025, doi: 10.24017/science.2025.2.8

Abstract: The rapid growth of computer-based technologies has transformed many sectors, with artificial intelligence playing a key role in automating tasks previously performed by humans. In this context, natural language processing models such as chatbots, including Chat Generative Pre-Trained Transformer (ChatGPT), are increasingly being used as analytical tools alongside traditional machine learning algorithms. However, despite these advancements, concerns remain regarding the accuracy, processing time, and overall reliability of ChatGPT compared to traditional coding-based machine learning algorithms. This study provides a comparative evaluation of ChatGPT's ability to generate intelligent responses. It focuses on three key aspects: accuracy across various datasets at different time intervals using the same account, performance relative to traditional machine learning algorithms in terms of accuracy, and the variability of ChatGPT's results across diverse data sources. To address these concerns, 15 algorithms were tested against ChatGPT. Tests were done at four different time intervals using healthcare and education datasets. ChatGPT showed competitive accuracy but had more variability and slower processing. As a result, this study highlights notable performance limitations for ChatGPT. For instance, in the heart disease dataset, the Random Forest model achieved an accuracy of 0.672 in 0.012 seconds, whereas the average performance of ChatGPT was 0.608 with a processing time of 0.274 seconds. In comparison, the traditional Gradient Boosting Machine model attained an accuracy of 0.623 in 0.124 seconds, while ChatGPT recorded an accuracy of 0.589 in 1.019 seconds. Finally, this study draws specific conclusions based on the results and offers recommendations for future research.

1. Introduction

The rapid progression of computer-based information technology has had a significant impact on how many facets of human life are changing. The most recent technical innovation brought about by the accelerated development of technology is artificial intelligence (AI) [1-3]. AI is a widely used technology in modern application development, as it allows computers to perform many tasks that humans can do. This facilitates individuals in addressing their diverse needs more effectively [4, 5]. Chat Generative Pre-Trained Transformer (ChatGPT) is a natural language processing (NLP) technology and AI language chatbot that is driven by GPT, which is a series of advanced language models created by

OpenAI [6]. These models are designed to understand and produce text that closely resembles human language. The core of GPT is built on the transformer architecture, which uses a technique called self-attention to efficiently process and generate text sequences, allowing it to handle complex language tasks with impressive accuracy [7], along with other cutting-edge technologies in the field of machine learning (ML) [8, 9]. It has drawn a lot of attention, especially on social media, for its amazing capacity to produce prose that appears human and the ability to have interactive discussions. NLP and ML approaches are combined in ChatGPT, a sophisticated language model, to create high-quality text that closely resembles human writing in a variety of languages and circumstances [10, 11]. When ChatGPT was released for Android on July 25, 2023, the OpenAI language model received a rating of 4.8 until the beginning of 2024 [12].

ChatGPT is capable of understanding the context of conversations, producing logical writing that mimics human speech, even though they are interacting with computers. However, the existence of TalkGPT also raises concerns about how this software may eventually replace human work [13]. Although the majority of reviews are positive, user reports indicate that ChatGPT gives incorrect responses, raising questions over the app's dependability [12]. Due to the increasing reliance on ChatGPT and similar AI tools, there is ongoing research aimed at addressing the various limitations, particularly related to processing time and response accuracy [14, 15]. However, the current literature lacks a clear and comprehensive understanding of these issues, as existing studies predominantly address specific limitations in isolation rather than offering an integrated evaluation [14]. More broadly, this work seeks to deepen our understanding of how intelligent systems based on AI and machine learning models perform in terms of accuracy and processing efficiency. A key gap in the existing research is the inconsistency in ChatGPT's responses and processing time when the same queries are repeated under identical account conditions at different times [16]. This study initially focuses on analyzing this variability, with further comparisons emphasizing its performance in the processing of large datasets and handling diverse feature sets.

This research is intended to address the following research questions (RQ):

RQ1: How accurate is ChatGPT for intelligent replying across different datasets, and how long does it take?

RQ2: How do traditional machine learning algorithms, implemented and analyzed through Python code, compare with ChatGPT's algorithm in terms of accuracy and processing time when generating intelligent responses based on the same datasets include the implementation steps for these traditional algorithms?

RQ3: How do ChatGPT outcomes across different datasets compare to intelligent answering?

The following is a summary of this paper's primary contributions:

- Conducted a comparative analysis of ChatGPT and established machine learning algorithms in terms of accuracy and processing time across benchmark datasets.
- Evaluated the consistency of ChatGPT's responses and processing times when identical queries were submitted at different time intervals.
- Identified the conditions under which ChatGPT's performance aligns with or deviates from that of traditional machine learning models.

The rest of the paper is structured as follows: Section 2 reviews related works, highlighting previous studies and the gaps this research aims to fill. In section 3, the materials and methods are discussed in detail, as well as the experimental setup. Section 4 presents the results of the experiments, looking at how the models performed based on various metrics. Section 5 delves into the interpretation and the discussion of these results. Finally, section 6 wraps up the paper with a summary of the key findings, in addition to suggestions for future research.

2. Related Works

AI and deep learning have significantly influenced diverse industries over the past decade, transforming communication, productivity, and problem-solving processes [17, 18].

Oghaz *et al*. [19] examined the detection and classification of ChatGPT-generated content using deep transformer models. Their study demonstrated that transformer-based approaches could reliably identify AI generated outputs, thereby supporting quality control and authenticity verification across various domains. Following these developments, Zhao *et al*. [20] introduced ChatAgri, exploring the potential of ChatGPT for cross-linguistic agricultural text classification. Their work highlighted the effectiveness of ChatGPT in domain-specific language tasks, particularly when adapted for multilingual agricultural datasets. Similarly, Tri Julianto *et al*. [21] proposed alternative text preprocessing strategies using ChatGPT, showing that AI-assisted preprocessing improved the efficiency and quality of subsequent NLP tasks.

Guo *et al*. [22] investigated the improvement path of the legal system concerning ChatGPT applications by integrating decision tree algorithms. Their results indicated that such an integration could enhance the transparency and explainability of AI-assisted decision-making processes. Ray [23] conducted a comprehensive review of ChatGPT, covering its background, applications, challenges, biases, ethical concerns, limitations, and future research directions, and concluded that while ChatGPT presents vast opportunities, addressing bias and ethical issues remains critical.

In terms of application domains, Haleem *et al*. [24] explored ChatGPT's competence in customer and patient service management. Their findings suggested that ChatGPT could effectively support real-time assistance and improve service quality. Koubaa *et al*. [25] conducted a broad survey of ChatGPT's capabilities and limitations, reporting that while the model excels in generating human-like responses, it still faces challenges in contextual understanding and factual accuracy.

Within specialized evaluation contexts, Kung *et al*. [26] assessed ChatGPT's performance on the United States Medical Licensing Examination. Their results demonstrated that the model achieved passing performance in several sections, suggesting its potential role in AI-assisted medical education. Azaria *et al*. [27] emphasized ChatGPT's value as a tool for experts, highlighting its strengths in augmenting expert decision-making while cautioning against overreliance due to its occasional inaccuracies.

Finally, Dwivedi *et a*l. [28] provided multidisciplinary perspectives on the implications of generative conversational AI, addressing its opportunities, challenges, and potential impact on research, practice, and policy. They stressed that while ChatGPT enables unprecedented automation and creativity, its limitations when it comes to contextual reasoning, bias mitigation, and robustness in varied input formats require further research attention.

Collectively, prior studies indicate that while ChatGPT represents a significant milestone in generative AI, continuous improvement is needed to address its limitations regarding context awareness, bias reduction, and adaptability to diverse applications [29].

To provide context and support the comparison, the relevant literature highlighting previous applications of machine learning algorithms to these datasets has been summarized in table 1.

Table 1: Overview of the previous works on dataset selection and evaluation in this study.

Ref	Year	Description	Dataset Name	Algorithm	Result
[30]	2022	This study evaluates machine learning algorithms for predicting student mental health issues, comparing K Nearest Neighbors (KNN) and Support Vector Machines (SVM) performance.	Student Mental Health, Healthcare	KNN, SVM	KNN achieved higher accuracy and but was very slow, while SVM delivered slightly lower accuracy yet much faster performance.
[31]	2023	This study compares various ML algorithms for diabetes prediction, achieving the highest accuracy.	Diabetes Prediction, Healthcare	Decision Trees (DT), Random Forest (RF), SVM	Achieved the highest accuracy of 96.26% in diabetes prediction.
[32]	2023	This systematic review explores the potential applications of ChatGPT in healthcare, focusing on the use of ML algorithms for healthcare data analysis.	Healthcare	Various ML	The study discusses the versatility of ML algorithms in healthcare applications using ChatGPT.

Table 1: Continue

[33]	2023	This paper compares the performance of Logistic Regression (LoR) and RF in the prediction of student qualifications based on performance metrics.	Education	LoR, RF	RF outperformed LoR in prediction accuracy.
[34]	2023	Explores the benefits and challenges of using ChatGPT in education, comparing the performance of multiple ML algorithms.	Education	KNN, SVM, RF, DT	SVM showed the best performance in educational data analysis using ChatGPT.
[35]	2024	Introduces the Multi-Role ChatGPT Framework (MRCF) to enhance performance in medical data analysis, focusing on accuracy, error reduction, and time efficiency.	Healthcare	Neural Networks (NN)	MRCF improved accuracy, reduced errors, and increased time. efficiency in healthcare data analysis.
[36]	2024	Presents a method for automated healthcare data classification using ML algorithms.	Healthcare	SVM, RF	accurate healthcare data classification with the tested ML algorithms.
[37]	2024	Uses educational data to analyze the performance of ML algorithms by ChatGPT.	Education	NN	Demonstrates the effectiveness of NN in educational data analysis.
[38]	2025	This study compares ML algorithms for heart disease detection, highlighting the accuracy of RF, KNN, and Principal Component Analysis (PCA).	Heart Disease Detection, Healthcare	KNN, Naive Bayes (NB), RF, PCA	RF and PCA achieved 99.4% accuracy in heart disease detection.
[39]	2025	Evaluates algorithms to predict student grades based on various metrics, including a stacking ensemble that outperforms individual models.	Student Performance, Education	KNN, RF, LoR,	Stacking ensemble outperformed individual models for student performance prediction.
[40]	2025	Surveys the accuracy of various ML algorithms in diagnosing heart disease, highlighting strengths and weaknesses.	Heart Disease Prediction, Healthcare	LoR, SVM, RF, XGBoost	XGBoost demonstrated the best performance in heart disease prediction.
[41]	2025	Applies PCA for dimensionality reduction, followed by a comparative study of various algorithms.	Healthcare	PCA, DT, KNN, XGBoost	XGBoost outperformed other algorithms in healthcare data analysis dataset.
[42]	2025	Evaluates KNN performance and compares with DT.	Healthcare	KNN, DT	KNN achieved 76% accuracy, with performance improvements using gridsearch.

2.1. Machine Learning and Algorithms

Since simpler models are unable to achieve the same level of performance, the complexity of novel ML models is a significant factor in the generation of results [43]. However, the opaqueness of popular black-box algorithms has raised issues as the models have grown. The models' dependability, equity, and accountability are some of the issues brought about by the lack of transparency [44]. ML algorithms are a collection of guidelines or procedures that an AI system uses to carry out tasks, usually to predict output values from a given set of input variables or to find new patterns and insights in the data. ML can learn, thanks to algorithms [45].

The present analysis integrates a diverse set of machine learning algorithms, each offering distinct advantages for tasks such as classification, regression, clustering, and dimensionality reduction. K Nearest Neighbors (KNN) is a simple yet effective algorithm that predicts labels based on the closest data points in the feature space [46]. Decision Tree (DT) build a tree-structured model by recursively splitting the data based on selected features [47]. Random Forest (RF) enhances this by aggregating multiple DTs, offering improved accuracy and robustness [48]. Support Vector Machines (SVM) classifies data by finding an optimal hyperplane that separates classes with the maximum margin [49]. Naive Bayes (NB) [50] applies probabilistic reasoning based on Bayes' theorem, assuming feature independence, and is particularly effective in text classification tasks [51]. Logistic Regression (LoR) [52]

models binary outcomes by estimating the probability of class membership using a logistic function [53], while Linear Regression (LiR) [31] captures the linear relationships between variables for continuous outcome prediction [54].

According to the specifications of this study, Principal Component Analysis (PCA) an unsupervised dimensionality reduction technique, was employed to enhance computational efficiency by reducing data complexity. To achieve meaningful results in terms of both accuracy and processing time, PCA was combined with the RF classifier, which ensured robust classification and reliable performance. PCA, as a statistical technique widely used in machine learning and data science, identifies directions known as principal components that capture the maximum variance in the data, and by projecting high-dimensional datasets onto a lower-dimensional space, it preserves most of the variability while simplifying analysis. PCA methods [55] transform high-dimensional data into a lower-dimensional space by identifying directions of maximum variance, whereas Linear Discriminant Analysis (LDA) [56] focuses on maximizing class separability. Ensemble methods such as Gradient Boosting Machines (GBM) [57] and AdaBoost (AB) [58] sequentially train weak learners to correct previous errors, creating a stronger overall model [59]. Hierarchical Clustering (HC) [60] organizes data into a nested hierarchy of clusters using agglomerative or divisive strategies, providing insights into data structure and similarity.

Neural Network (NN) [61], including architectures such as Convolutional Neural Networks, Recurrent Neural Networks, and transformers, mimic brain-like structures to model complex, non-linear patterns across diverse tasks such as image and speech recognition. Ridge Regression (RR) [62] extends LiR by introducing regularization to handle multicollinearity and prevent overfitting. Multilayer Perceptrons (MLP) [63], a type of deep feedforward NN, utilize multiple hidden layers with nonlinear activation functions to capture complex relationships in structured or unstructured data.

3. Materials and Methods

To clarify the methodology, the chart below provides a step-by-step illustration of the research process. It begins with the dataset selection and experimental setup, proceeds through the ChatGPT and Python-based analyses, and concludes with the statistical evaluation, comparison, and identification of the optimal time and accuracy for processing. Seven steps were used as the research stages in this study. These phases are illustrated as shown in figure 1.

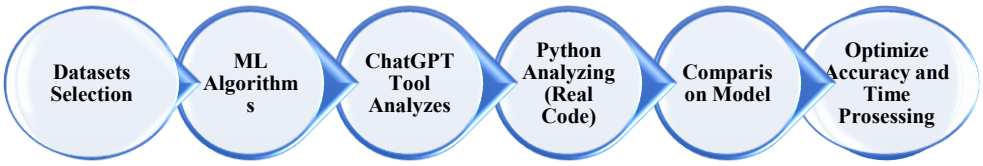


Figure 1: Six sequential processes as steps of methodology research.

3.1. Datasets

The method for assessing processing time and accuracy field consisted of open-source datasets used to compare and assess the results at four distinct times. The research datasets were acquired from Kaggle, a popular website that offers a range of datasets for application in machine learning and data science. This study made use of four datasets, as shown in table 2.

Table 2: Overview of the selected datasets used in this study.

No	Ref	Name	Dataset Type	No. of Col.	No. of Rows	URL
1	[64]	Heart Disease (HD)	Health	16	920	https://www.kaggle.com/datasets/redwankarim-sony/heart-disease-data
2	[65]	Disease Sign and Symptom (DSS)	Health	10	349	https://www.kaggle.com/datasets/uom190346a/dis-ease-symptoms-and-patient-profile-dataset
3	[66]	Student Perfor-mance (SP)	Educa-tion	15	2392	https://www.kaggle.com/datasets/rabieelkha-roua/students-performance-dataset
4	[67]	Student World University (SWU)	Educa-tion	33	649	https://www.kaggle.com/datasets/larsen0966/stu-dent-performance-data-set

3.2. Experimental Setup

First, the data was downloaded from the Kaggle website. To achieve and demonstrate the research objectives, four experiments were conducted utilizing ChatGPT. Four experiments were then performed using traditional machine learning code implemented in Python, which is a high-level, interpreted programming language that is widely used for data analysis, machine learning, and automation. Here it was employed to implement and evaluate the proposed approach [68]. This test illustrates the comparison between the results generated by ChatGPT and those obtained through actual code execution. It also highlights how to determine the most effective time to use the ChatGPT analysis tool. The findings demonstrate clear differences in both results and time complexity across the different testing times, even within the same time zone. The TMLC means using traditional machine learning models implemented in Python code, providing the baseline for comparison with the ChatGPT-generated outputs. Both experiments were applied for all 15 machine learning algorithms.

ChatGPT was used to enhance the ML algorithms at various points to attain the highest accuracy. Below is the list of the common pure ML algorithms without K-fold and any modifications that were analyzed for accuracy and time processing in this research. Accordingly, to evaluate the accuracy and processing time (in seconds) of the different machine learning algorithms, specific parameters were used for each model to ensure fair and effective testing, as illustrated in table 3.

Table 3: Machine learning algorithms used for testing and determining accuracy and time processing (Sec.).

No.	Algorithms	Parameters
1	K Nearest Neighbors	n-neighbours: 5, weights: uniform, leaf_size:30, test:0.2, random state:40
2	Decision Trees	criterion: gini, max_depth: none, min_samples_split: 2, min_samples_ leaf: 1.
3	Random Forest	n-estimators: 100, random state:40 test: 0.2.
4	Support Vector Machine	kernel: RBF, C: 1.0, gamma: scale, random state: 40.
5	Naive Bayes	Var-smoothing: (1e-9)
6	Logistic Regression	Random state: 40, test size:0.2
7	Linear Regression	Fit-intercept: true, normalize: false, test size: 0.2, n_job: 1, random state: 40.
8	Multilayer Perceptron	hidden-layer-sizes: (100, 50), max_iter: 1000, random state:40.
9	Ridge Regression	alpha: 1.0, random state: 40.
10	Gradient Boosting Machines	n-estimators: 100, max_depth: 3, subsample: 1.0, learning rate: 0.1.
11	AdaBoost	n-estimators: 50, learning-rate: 1.0, random state:40.
12	Hierarchical Clustering	linkage: ward, affinity: euclidean, compute_full_tree: auto.
13	Neural Networks	hidden-layer-sizes: (100, 50), max_iter: 1000, random state:40.
14	Linear Discriminant Analysis	solver:std, priors:none, test size:0.2, random state:40.
15	Principal Component Analysis	n-components: 0.95, random state: 40

According to the study methodology, the results were derived using the ChatGPT tool, with all applications implemented through Python code. This investigation has outlined the necessary steps to initiate the testing process within two comparative strategies. Each strategy has been thoroughly explained, accompanied by visual representations of the problem to enhance understanding and illustrate the evaluation process.

In the first strategy, a single user account was created on the ChatGPT platform, followed by uploading the dataset intended for analysis regarding processing time and accuracy. Next, the operational rules for ChatGPT were configured in accordance with the experimental requirements. If the dataset was incompatible with the algorithm under examination and yielded no valid output, the result was recorded as "no result." Conversely, if the dataset was compatible, then accuracy and processing time were calculated before proceeding to the final stage. The analysis for each algorithm was then repeated four times at 12:00 a.m., 6:00 a.m., 12:00 p.m., and 6:00 p.m. Once the first algorithm had been analyzed, the same procedure was applied to the remaining algorithms using the same dataset. Finally, all of the aforementioned steps were repeated for the three additional datasets, each tested with the same 15 algorithms, as illustrated in figure 2.

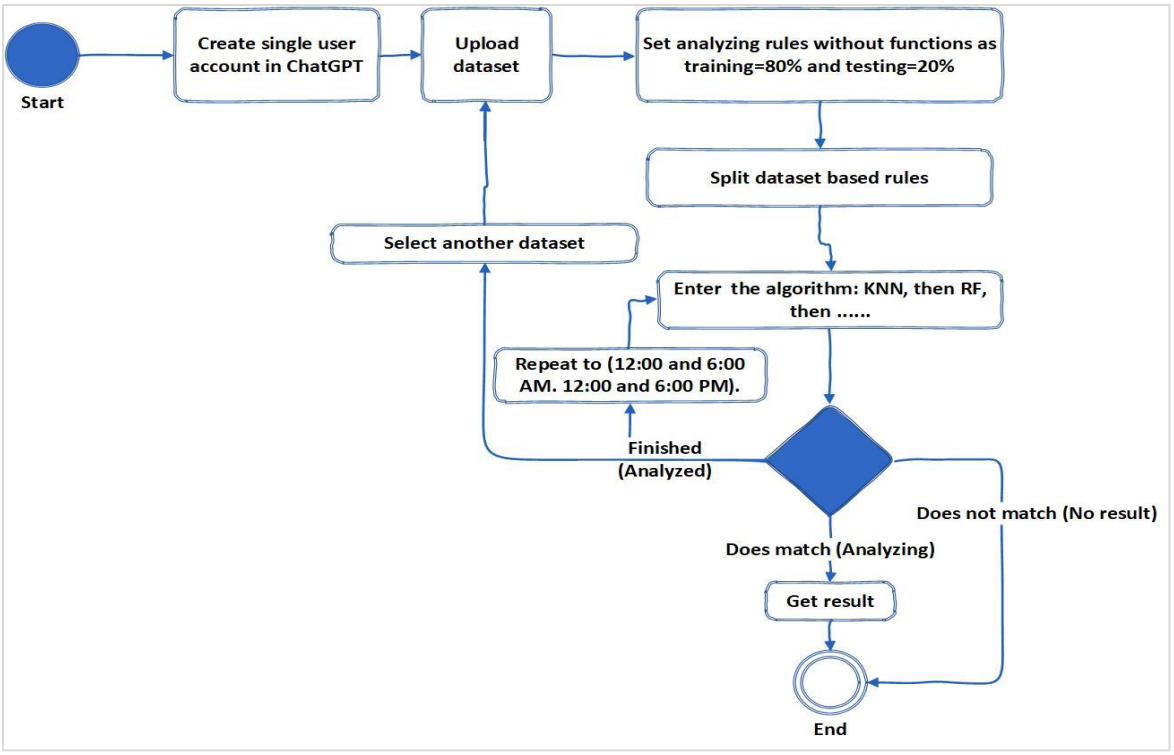


Figure 2: Steps for determining accuracy and time processing (Sec.) for the 15 algorithms using ChatGPT.

The second strategy was dedicated to demonstrating the evaluation of Jupyter, the Python programming language and Libraries, which were the three main software and tools used to prepare the offline environment in which to conduct the analysis. Jupyter Notebook is an open-source software that allows users to create and share documents containing live code, equations, visualizations, and narrative text. While Jupyter itself is a platform, it is primarily used with Python, and the term "Jupyter Python library" often refers to the Python libraries commonly used within a Jupyter Notebook environment [69].

The analysis of the 15 algorithms of ML to determine the accuracy and time processing (sec.) of each was done using Python on a local computer without access to the Internet. The applications were run once because the results at different times do not change. The local computer was prepared for analysis using the following steps: the first step was to install the Jupyter software and prepare the required libraries. In the second step, a dataset was uploaded for the analysis of the accuracy and

processing time (sec.). The third step involved handling missing values and preparing the dataset for further analysis. In the fourth step, the rules were tested in practice. The fifth step was to proceed to the final stage if the dataset matched the algorithm for analysis. However, if the dataset was incompatible and resulted in failure, it was disregarded, and the output recorded as *"No result"*. In the sixth step, each procedure was repeated four times at 12:00 AM, 6:00 AM, 12:00 PM, and 6:00 PM. Once the first algorithm had been processed, the same procedure was repeated for the other algorithms using the same dataset. Finally, the steps were repeated for three additional datasets, each tested with the same 15 algorithms, as illustrated in figure 3.

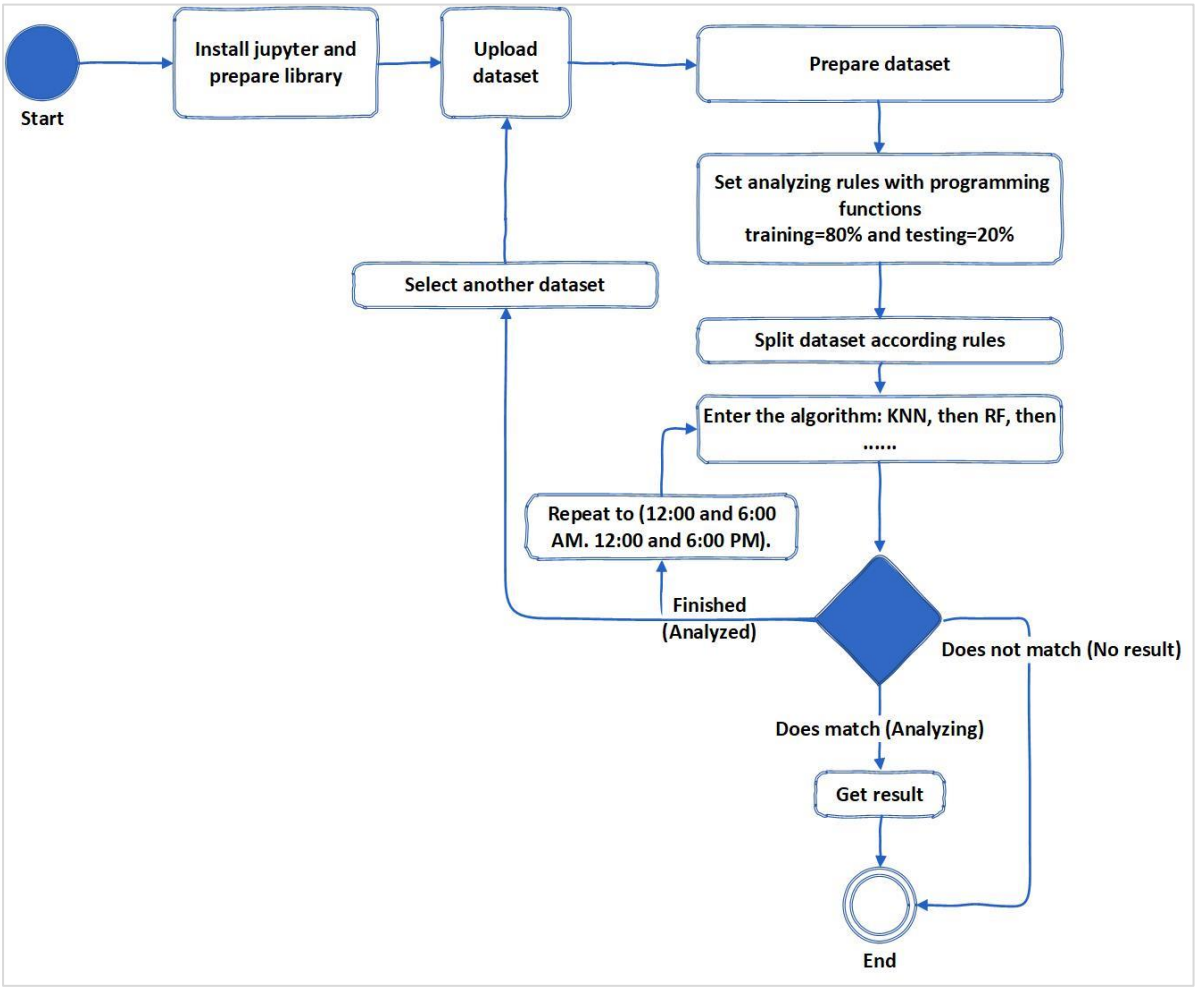


Figure 3: Steps for finding accuracy and time processing (Sec.) of the 15 algorithms using Python programming.

Five comparisons were presented to show the best accuracy and time processing as follows:

1. Determining and selecting the best accuracy and time processing for the health (heart disease with disease sign and symptom) dataset.
2. Determining and selecting the best accuracy and time processing for education (student performance with student world university).
3. Comparison and evaluation of the health results with the Python programming test results.
4. Comparison and evaluation of the education results with the Python programming test results.
5. Comparison and evaluation between the health result and education for time processing.

3.3. Statistical Analysis

The average of the provided available and unavailable values was determined for each component of the accuracy and processing time results. To compare the accuracy and time processing outcomes for the four distinct datasets, the maximum and minimum values were also determined. IBM SPSS Statistics V.30 was used for the statistical analysis.

The best accuracy and time processing were found by calculating the mean average (AVG), maximum, and minimum. equation (1) tells us the accuracy and time processing, calculated by dividing the total accuracy and time processing over 16 days. While the minimum average of time processing is the best time processing, the maximum of accuracy is the highest accuracy according to equations (2, 3) from each round for accuracy.

$$A = \frac{1}{n}\sum_{i=1}^{n} ai \tag{1}$$

$$\text{Maximum} = \text{Max}\ (N_1, N_2\ , \ldots\ N_{last}) \tag{2}$$

$$\text{Minimum} = \text{Min}\ (N_1, N_2\ , \ldots\ N_{last}) \tag{3}$$

For descriptive statistics, standard deviation (STD) was employed as shown in equation (4). The relationship between accuracy and time processing was investigated using univariate analysis, as shown in equations (5, 6). When the P value was less than 0.05, it was considered statistically significant.

$$STD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^{\,2}} \tag{4}$$

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{s_1}^2}{n_1}} + \sqrt{\frac{{s_2}^2}{n_2}}} \tag{5}$$

$$P - value = P(T \geq |t|) \tag{6}$$

In the context of equations (4, 5, 6), n represents the number of observations or values in a given sample, while xi denotes each individual value within that sample. The term $\bar{x}$ refers to the sample mean, which is the average of all xi values. When comparing two samples, $s1$ and $s2$ represent the variances of the first and second samples, respectively, while $n1$ and $n2$ denote the corresponding sample sizes. These components are used to compute the t-statistic, which measures the difference between the two sample means relative to the variability in the data, and the resulting p-value indicates the statistical significance of that difference.

4. Results

Fifteen algorithms used to analyze the best accuracy and time processing were included in this study. The results were classified according to the ChatGPT and Python programming results, specifically accuracy, time processing and value availability. The maximum value of accuracy was the best accuracy, while the minimum value of time processing was the best time processing. The response availability of each algorithm was a different percentage time.

4.1. ChatGPT Results

This section presents the results obtained from multiple testing sessions, highlighting comparative performance across four different time periods using all selected datasets. The analysis emphasizes that the ChatGPT tool played a significant role in generating outcomes during each session, revealing noticeable variations in time complexity over different runs. As noted, ChatGPT functions not only as a

chatbot for answering questions but also holds potential as a machine learning assistant, capable of guiding users in model development and learning processes without directly providing code.

The results demonstrate how model performance varied over time. This framework tested several models to determine the most effective configuration based on different time intervals. Some trials experienced error rates or incomplete results, which were attributed to ChatGPT's processing limitations during the model execution. Each dataset was tested four times, and comparisons were made within the same dataset across different time runs. It is important to note that the datasets were not compared against each other as the purpose of this analysis was to evaluate consistency and variation within the individual datasets across selected timeframes.

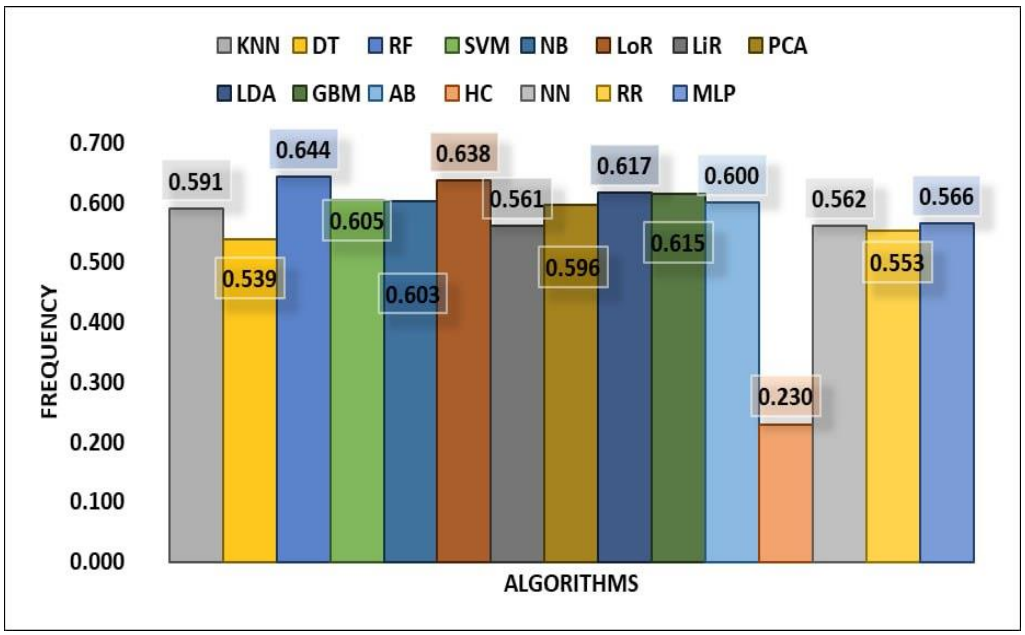
4.1.1. Heart Disease Dataset

HD is one of the healthcare datasets used to confirm the performance of ChatGPT's accuracy and time limitation. The RF algorithm had the highest accuracy of 0.644, while the RR algorithm had the best time processing of 0.003. The best accuracy for each of the algorithms (KNN, DT, RF and SVM) was (0.591, 0.539, 0.644 and 0.605), and the best processing time for the same algorithm was (0.011, 0.005,0.231 and 0.42). More details about the HD accuracy and time processing were recorded, The STD values indicate the consistency of each algorithm's performance. Lower STD values reflect greater stability across the time intervals, while higher values suggest variability. For example, LDA demonstrates high consistency, whereas HC shows significant fluctuations, as shown in table 4 and figure 4.

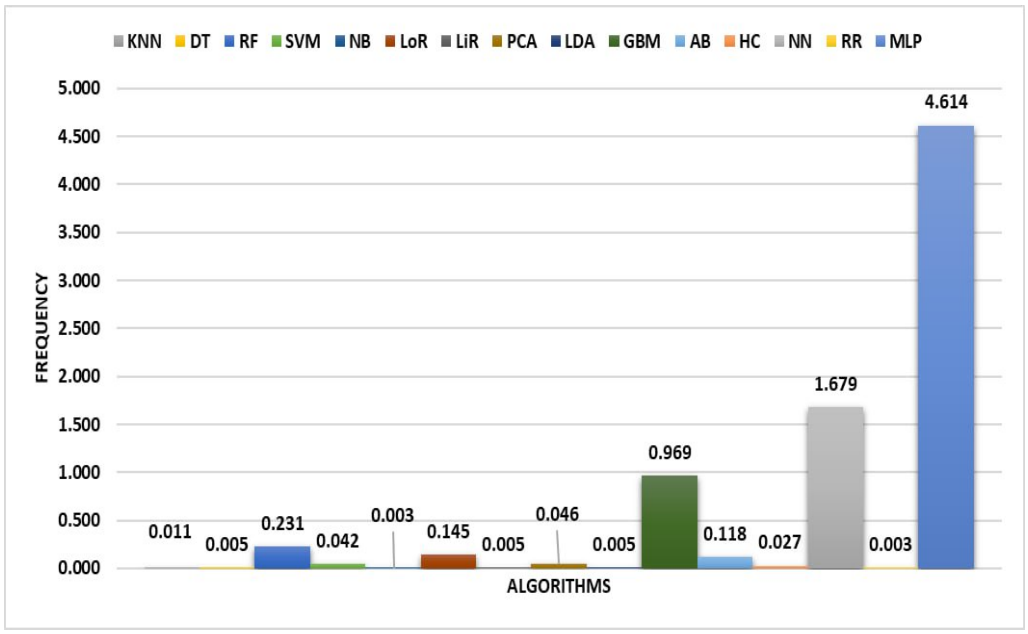
Table 4: Best results for accuracy and time processing (Sec.) for the HD dataset at different times according to the AVG, STD and P-values.

Algorithms	12:00 AM			6:00 AM			12:00 PM			6:00 PM			P-Value
	Accuracy		Time	Accuracy		Time	Accuracy		Time	Accuracy		Time	
	AVG	STD		AVG	STD		AVG	STD		AVG	STD		
KNN	0.565	0.039	**0.011**	0.548	0.052	0.075	0.559	0.036	0.013	**0.591**	0.087	0.041	0.0001*
DT	0.511	0.045	0.006	0.464	0.028	**0.005**	0.488	0.035	0.008	**0.539**	0.104	0.006	0.0003*
RF	0.592	0.033	0.287	0.598	0.023	**0.231**	0.596	0.016	0.299	**0.644**	0.086	0.281	0.0003*
SVM	0.584	0.058	0.043	0.564	0.071	**0.042**	0.569	0.050	0.044	**0.605**	0.090	0.046	0.0000*
NB	0.552	0.034	0.004	0.512	0.100	**0.003**	0.550	0.027	0.005	**0.603**	0.099	0.004	0.0000*
LoR	0.597	0.019	**0.145**	0.606	0.014	0.214	0.592	0.019	0.223	**0.638**	0.094	0.307	0.0006*
LiR	0.487	0.056	0.010	0.468	0.064	**0.005**	0.486	0.025	0.011	**0.561**	0.164	0.008	0.0001*
MLP	0.554	0.034	6.384	**0.566**	0.040	14.90	0.555	0.032	5.349	0.559	0.057	**4.614**	**0.0560**
RR	0.535	0.052	0.009	0.517	0.110	**0.003**	**0.553**	0.063	0.008	0.550	0.109	0.006	0.0000*
GBM	0.581	0.051	1.053	0.593	0.018	1.020	0.569	0.043	1.033	**0.615**	0.079	**0.969**	**0.0005***
AB	0.577	0.032	0.128	0.560	0.035	**0.118**	0.559	0.026	0.127	**0.600**	0.086	0.134	0.0000*
HC	0.205	0.116	**0.027**	N/A	N/A!	N/A	0.160	0.127	0.067	**0.230**	0.265	0.038	0.0000*
NN	0.559	0.031	4.414	**0.562**	0.028	2.682	0.559	0.030	**1.679**	0.561	0.055	1.928	0.0417*
LDA	0.580	0.018	0.007	0.590	0.016	**0.005**	0.588	0.009	0.007	**0.617**	0.086	0.010	0.0000*
PCA	0.563	0.050	**0.046**	**0.596**	0.016	0.247	0.575	0.029	0.134	0.571	0.072	0.185	0.0013*

* Statistically Significant



A- Accuracy



B- Time Processing

Figure 4: Best results for (A) accuracy and (B) time processing (Sec.) for the HD dataset at different times.

In the HD dataset analysis, the accuracy availability for algorithms (KNN, DT, RF, SVM, NB and LoR) was 100% during the twenty-four hour test period. The HC algorithm had the worst availability compared to other algorithms, and had a result of 50% at 12:00 PM. The largest non-availability was recorded for the same algorithm, which was 100% at 6:00 AM, as shown in table 5. Unavailable data indicates variations in response rates (Truth Rate) based on time zone during the analysis. Algorithm availability refers to the ability to produce responses at specific times (6:00 and 12:00 AM, 6:00 and 12:00 PM), which helps determine response errors and error rates. For example, KNN maintained a 100% response rate at all times, while HC responded 50% of the time at 12:00 PM and failed to respond entirely at 6:00 PM.

Table 4: Best available and no available values for the HD dataset at different times.

No.	Algorithms	Available		No Available (Error Rate)	
		Availability	%	Availability	%
1	KNN	24 Hours	100	24 Hours	0
2	DT	24 Hours	100	24 Hours	0
3	RF	24 Hours	100	24 Hours	0
4	SVM	24 Hours	100	24 Hours	0
5	NB	24 Hours	100	24 Hours	0
6	*LoR*	24 Hours	100	24 Hours	0
7	*LiR*	6:00 PM	81.25	6:00 AM	81.25
8	MLP	6:00 AM	93.75	12:00 AM, 12:00 PM & 6:00 PM	12.5
9	RR	6:00 PM	91.25	6:00 AM	75
10	GBM	12:00 AM, 6:00 AM & 6:00 PM	100	12:00 PM	6.25
11	AB	12:00 AM, 12:00 PM & 6:00 PM	100	6:00 AM	6.25
12	HC	12:00 PM	50	6:00 AM	100
13	NN	6:00 AM & 6:00 PM	93.75	12:00 AM & 12:00 PM	6.25
14	LDA	12:00 AM & 6:00 PM	100	6:00 AM	25
15	PCA	12:00 PM	62.5	6:00 AM	75

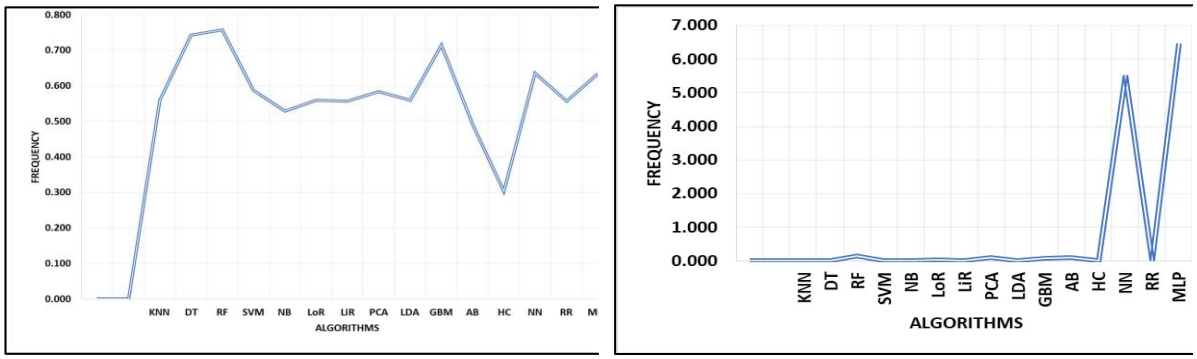
4.1.2. Disease Sign and Symptom Dataset

One of the health care datasets used to verify ChatGPT's accuracy and time limit is the DSS, where the RF algorithm had the highest accuracy of 0.757, while the NB and LDA algorithms had the best time processing of 0.003. The best accuracy for each of the algorithms (KNN, DT, RF and SVM) was (0.560, 0.743, 0.757 and 0.588), and the best processing time for the same algorithm was (0.007, 0.004,0.161 and 0.010). More details about DSS accuracy and time processing were recorded. The STD values in this table highlight the performance stability of each algorithm. Lower STD values suggest consistent outcomes across different times, as seen in models like NB and LDA. In contrast, HC and RR exhibit higher STD values, as shown in table 6 and figure 5.

Table 5: Best results for accuracy and time processing (Sec.) for the DSS dataset at different times according to the AVG, STD and P-values.

Algorithms	12:00 AM			6:00 AM			12:00 PM			6:00 PM			P-Value
	Accuracy		Time	Accuracy		Time	Accuracy		Time	Accuracy		Time	
	AVG	STD		AVG	STD		AVG	STD		AVG	STD		
KNN	0.549	0.040	0.008	0.554	0.045	0.007	0.547	0.040	0.009	0.560	0.048	0.008	0.0000*
DT	0.725	0.072	0.005	0.741	0.007	0.004	0.742	0.004	0.005	0.743	0.000	0.006	0.0000*
RF	0.746	0.044	0.231	0.754	0.008	0.161	0.754	0.011	0.201	0.757	0.000	0.193	0.0000*
SVM	0.583	0.023	0.013	0.587	0.021	0.019	0.580	0.019	0.011	0.588	0.025	0.010	0.0000*
NB	0.528	0.002	0.004	0.529	0.000	0.003	0.529	0.000	0.004	0.529	0.000	0.004	0.0000*
LoR	0.559	0.006	0.042	0.557	0.000	0.025	0.557	0.000	0.046	0.557	0.000	0.038	0.0000*
LiR	0.557	0.000	0.014	0.557	0.000	0.005	0.526	0.097	0.008	0.557	0.000	0.007	0.0000*
MLP	0.622	0.052	6.716	0.631	0.052	7.587	0.628	0.051	6.453	0.634	0.055	6.515	0.0001*
RR	0.518	0.142	0.021	0.557	0.000	0.005	0.526	0.097	0.007	0.557	0.000	0.021	0.0000*
GBM	0.705	0.037	0.149	0.714	0.000	0.079	0.714	0.000	0.098	0.714	0.000	0.092	0.0000*
AB	0.490	0.016	0.112	0.491	0.015	0.100	0.486	0.000	0.114	0.486	0.000	0.116	0.0000*
HC	0.394	0.243	0.049	0.407	0.111	0.016	0.303	0.251	0.045	0.408	0.177	1.352	0.9713
NN	0.624	0.055	7.134	0.630	0.052	6.757	0.615	0.044	5.512	0.636	0.057	7.257	0.0006*
LDA	0.559	0.006	0.005	0.557	0.000	0.003	0.557	0.000	0.005	0.557	0.000	0.004	0.0000*
PCA	0.546	0.021	0.015	0.549	0.016	0.011	0.562	0.017	0.010	0.584	0.046	0.014	0.0000*

*Statistically Significant



A- Accuracy **B- Time Processing**

Figure 5: Best results for accuracy and time processing (Sec.) for the DSS dataset at different times.

There are both available and unavailable values in the DSS dataset, and these depend on the processing time and accuracy analysis time. Specifically, 11 algorithms have a 100% availability rate, with only four algorithms having unavailable values, as shown in table 7.

Table 6: Best available and no available values in the DSS dataset at different times.

No.	Algorithm	Available		No Available (Error Rate)	
		Availability	**%**	**Availability**	**%**
1	KNN	24 Hours	100	24 Hours	0
2	DT	24 Hours	100	24 Hours	0
3	RF	24 Hours	100	24 Hours	0
4	SVM	24 Hours	100	24 Hours	0
5	NB	24 Hours	100	24 Hours	0
6	*LoR*	24 Hours	100	24 Hours	0
7	*LiR*	12:00 PM	62.5	6:00 PM	68.75
8	MLP	24 Hours	100	24 Hours	0
9	RR	12:00 AM	81.25	6:00 AM	56.25
10	GBM	24 Hours	100	24 Hours	0
11	AB	24 Hours	100	24 Hours	0
12	HC	6:PM	31.25	6:00 PM	87.5
13	NN	12:00 AM, 6:00 AM & 12:00 PM	100	6:00 PM	6.25
14	LDA	6:00 AM & 6:00 PM	100	12:00 AM & 12:00 PM	6.25
15	PCA	6:00 PM	43.75	12:00 AM	75

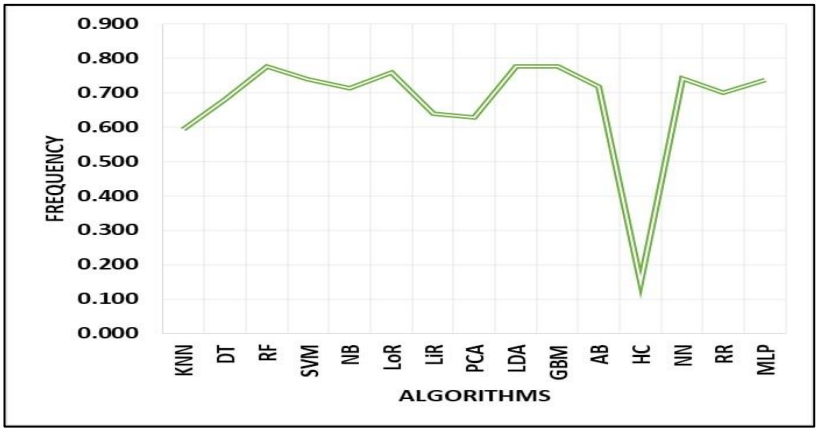
4.1.3. Student Performance Dataset

One of the educational datasets used to verify ChatGPT's accuracy and time restriction was the SP dataset. The RF and GBM algorithms had the highest accuracy at 0.777, while the NB algorithm had the highest time processing value of 0.004. The best accuracy for each of the algorithms (KNN, DT, RF and SVM) was (0.594, 0.680, 0.777 and 0.739) but the best processing time for the same algorithm was (0.034, 0.011,0.454 and 0.203). More details about the SP's accuracy and time processing were recorded, Algorithms such as LDA and NB demonstrate low STD values, indicating consistent results across different time intervals. In contrast, models like RR and HC exhibit high variability, as shown in table 8 and figure 6.

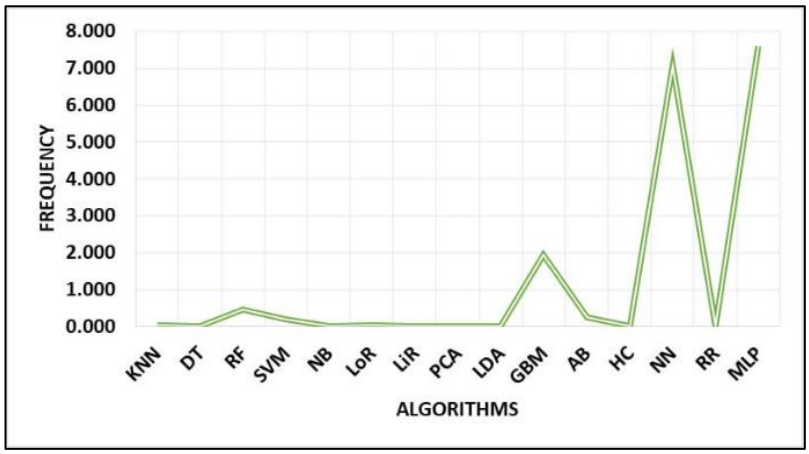
Table 7: Best result of accuracy and time processing (Sec.) of SP dataset at different time according to the AVG, STD and P-value.

Algorithms	12:00 AM			6:00 AM			12:00 PM			6:00 PM			P-Value
	Accuracy		Time	Accuracy		Time	Accuracy		Time	Accuracy		Time	
	AVG	STD		AVG	STD		AVG	STD		AVG	STD		
KNN	0.593	0.047	0.036	0.594	0.048	0.034	0.588	0.045	0.041	0.575	0.029	0.041	0.0000*
DT	0.680	0.112	0.013	0.639	0.078	0.011	0.653	0.084	0.013	0.649	0.085	0.012	0.0000*
RF	0.777	0.082	0.512	0.744	0.047	0.454	0.755	0.062	0.589	0.753	0.063	0.531	0.0031*
SVM	0.739	0.043	0.235	0.727	0.026	0.203	0.725	0.032	0.252	0.724	0.032	0.258	0.0000*
NB	0.713	0.056	0.004	0.688	0.031	0.004	0.697	0.043	0.005	0.696	0.044	0.005	0.0000*
LoR	0.758	0.041	1.116	0.741	0.021	0.138	0.747	0.031	0.153	0.746	0.033	0.031	0.2187
LiR	0.343	0.295	0.012	0.399	0.433	0.006	0.640	0.368	0.017	0.498	0.478	0.155	0.0084*
MLP	0.733	0.051	7.601	0.681	0.264	11.241	0.737	0.040	9.556	0.736	0.053	15.894	0.0099*
RR	0.463	0.263	0.029	0.040	0.000	0.005	0.700	0.128	0.021	0.677	0.121	0.016	0.0567
GBM	0.777	0.087	2.138	0.744	0.052	1.933	0.756	0.068	2.186	0.753	0.066	2.151	0.0001*
AB	0.717	0.107	0.244	0.656	0.002	0.249	0.701	0.097	0.251	0.716	0.107	0.263	0.0000*
HC	0.072	0.069	0.121	0.150	0.106	0.022	0.076	0.066	0.085	0.050	0.063	0.049	0.6691
NN	0.742	0.043	7.273	0.720	0.018	9.841	0.733	0.034	7.077	0.738	0.049	8.982	0.0015*
LDA	0.757	0.024	0.022	0.775	0.076	0.013	0.750	0.018	0.025	0.748	0.021	0.035	0.0000*
PCA	0.372	0.115	0.179	0.628	0.444	0.010	0.498	0.157	0.078	0.461	0.241	0.048	0.0179*

* Statistically Significant



A-Accuracy



B- Time processing

Figure 6: Best results for accuracy and time processing (Sec.) of the SP dataset at different times.

In the SP dataset analysis, the accuracy availability for the algorithms (KNN, DT, RF, SVM and NB) was 100% during the twenty-four-hour test period, but the PCA and HC algorithms had the worst availability compared to other algorithms, at 37.5% at 6:00 AM. The largest no-availability was recorded for the PCA and HC algorithms of 93.75% at 12:00 PM and 6:00 PM, as shown in table 9.

Table 8: Best available and no available values for the SP dataset at different times.

No.	Algo-rithms	Available		No Available (Error Rate)	
		Availability	%	Availability	%
1	KNN	24 Hours	100	/	0
2	DT	24 Hours	100	/	0
3	RF	24 Hours	100	/	0
4	SVM	24 Hours	100	/	0
5	NB	24 Hours	100	/	0
6	LoR	12:00 AM, 6:00 AM, 12:00 PM	100	6:00 PM	6.25
7	LiR	6:00 AM	56.25	12:00 PM & 6:00 PM	68.75
8	MLP	12:00 AM, 6:00 AM, 12:00 PM	100	6:00 PM	6.25
9	RR	12:00 AM & 6:00 PM	50	12:00 PM	62.5
10	GBM	12:00 AM, 6:00 AM, 6:00 PM	100	12:00 PM	6.25
11	AB	24 Hours	100	/	0
12	HC	12:00 PM & 6:00 PM	37.5	6:00 AM	93.75
13	NN	12:00 AM, 6:00 AM &6:00 PM	100	12:00 PM	6.25
14	LDA	6:00 AM	100	12:00 AM, 12:00 PM, 6:00 PM	6.25
15	PCA	12:00 PM, 6:00 PM	37.5	6:00 AM	93.75

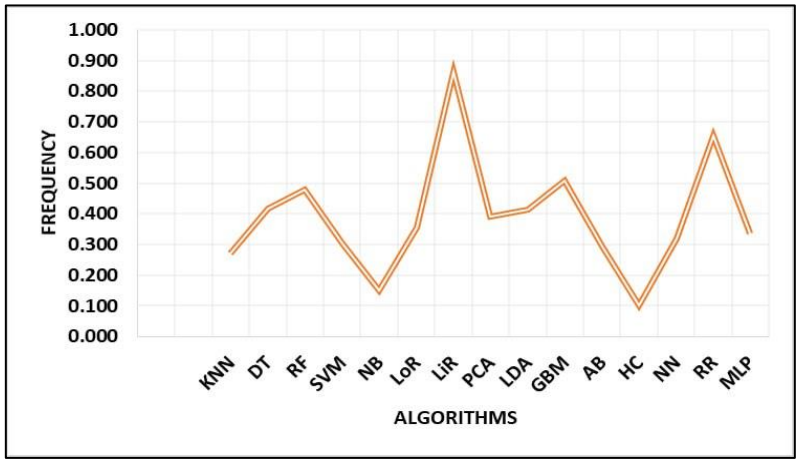
4.1.4. Student World University

The SWU dataset is one of the educational datasets that was used to verify ChatGPT's accuracy and time limit performance. The LiR algorithm had the highest accuracy of 0.862 but the NB algorithm had the best time processing at 0.005. The best accuracy for each of the algorithms (KNN, DT, RF and SVM) was (0.270, 0.418, 0.478 and 0.303), but the best processing time for the same algorithms was (0.170, 0.009,0.287 and 0.040). More details about the SWU accuracy and time processing were recorded, as shown in table 10 and figure 7.

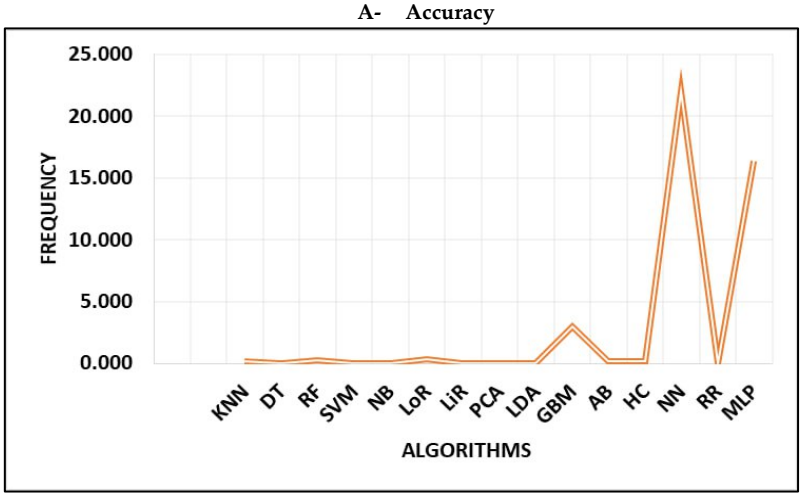
Table 9: Best result of accuracy and time processing (Sec.) of SWU dataset at different time according to the AVG, STD and P-value.

	12:00 AM			6:00 AM			12:00 PM			6:00 PM			P-Value
	Accuracy		Time	Accuracy		Time	Accuracy		Time	Accuracy		Time	
	AVG	STD		AVG	STD		AVG	STD		AVG	STD		
KNN	0.253	0.116	0.253	0.270	0.087	0.170	0.231	0.081	0.313	0.243	0.083	0.279	0.9153
DT	0.418	0.058	0.009	0.401	0.011	0.017	0.407	0.016	0.012	0.410	0.008	0.023	0.0000*
RF	0.478	0.069	0.296	0.466	0.018	0.330	0.463	0.018	0.287	0.468	0.008	0.287	0.0005*
SVM	0.294	0.123	0.069	0.303	0.060	0.040	0.261	0.054	0.052	0.267	0.057	0.050	0.0003*
NB	0.150	0.143	0.006	0.135	0.020	0.005	0.117	0.023	0.006	0.117	0.023	0.007	0.0005*
LoR	0.354	0.103	1.974	0.343	0.026	0.676	0.329	0.011	1.165	0.328	0.011	0.294	0.1502
LiR	0.575	0.334	0.010	0.862	0.000	0.007	0.559	0.207	0.014	0.680	0.226	0.013	0.0026*
MLP	0.301	0.119	24.122	0.335	0.136	25.138	0.333	0.164	16.340	0.321	0.126	20.002	0.0018*
RR	0.629	0.305	0.008	0.641	0.312	0.008	0.653	0.229	0.018	0.646	0.281	0.006	0.0000*
GBM	0.507	0.060	3.289	0.493	0.003	3.003	0.494	0.003	3.562	0.492	0.000	3.316	0.0001*
AB	0.293	0.097	0.150	0.269	0.000	0.181	0.269	0.000	0.150	0.269	0.000	0.136	0.0021*
HC	0.025	0.007	0.020	0.100	0.000	0.015	0.052	0.035	0.056	0.083	0.024	0.026	0.1908
NN	0.283	0.037	28.687	0.318	0.029	24.905	0.309	0.115	26.317	0.290	0.035	21.998	0.0003*
LDA	0.415	0.088	0.033	0.387	0.011	0.083	0.391	0.004	0.034	0.392	0.000	0.024	0.0002*
PCA	0.321	0.110	0.219	0.390	0.040	0.340	0.288	0.114	0.195	0.162	0.000	0.010	0.0179*

* Statistically Significant



A- Accuracy



B- Time Processing

Figure 7: Best results for accuracy and time processing (Sec.) in the SWU dataset at different times.

In the SWU dataset analysis, the accuracy availability for algorithms KNN, DT, RF, SVM, NB and LoR was 100% during the 24-hour test period, while the HC algorithm had the worst availability compared to other algorithms, at 31.25% at 12:00 PM. The largest no-availability was recorded for the same algorithm, which was 93.75% at 6:00 AM, as shown in table 11.

Table 10: Best available and no available values in the SWU dataset at different times.

No.	Algorithms	Available		No Available (Error Rate)	
		Availability	%	Availability	%
1	KNN	24 Hours	100	24 Hours	0
2	DT	24 Hours	100	24 Hours	0
3	RF	24 Hours	100	24 Hours	0
4	SVM	24 Hours	100	24 Hours	0
5	NB	24 Hours	100	24 Hours	0
6	LoR	24 Hours	100	24 Hours	0
7	LiR	12:00 PM & 6:00 PM	43.75	6:00 AM	68.75
8	MLP	24 Hours	100	24 Hours	0
9	RR	6:00 PM	43.75	6:00 AM	81.25
10	GBM	24 Hours	100	24 Hours	0
11	AB	12:00 AM, 6:00 AM, 6:00PM	100	12:00 PM	6.25
12	HC	12:00 PM	31.25	6:00 AM	93.75
13	NN	6:00 AM	100	12:00 AM, 12:00 PM & 6:00 PM	6.25
14	LDA	12:00 AM & 6:00PM	100	6:00 AM	87.5
15	PCA	12:00 PM	50	6:00 PM	87.5

Python Programming Results

During the evaluation of the four datasets within the proposed framework, the goal was to identify the most optimal testing time and determine the most efficient time complexity using ChatGPT tools. The final comparison focused on both accuracy and time complexity, aiming to highlight which dataset and model combination performed best. This also demonstrates the potential of using ChatGPT as a reliable tool during model testing and evaluation.

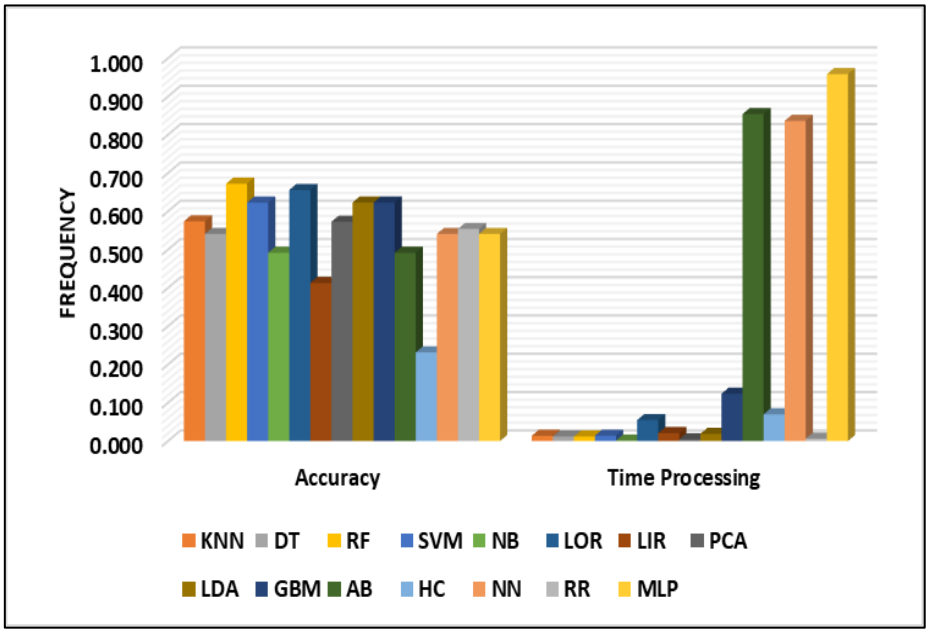
Although the Python programming results were used in the experiments, the models varied in structure and modality. Each dataset was tested accordingly, and the results will be discussed in detail in the following section.

4.2.1. Healthcare Dataset Results

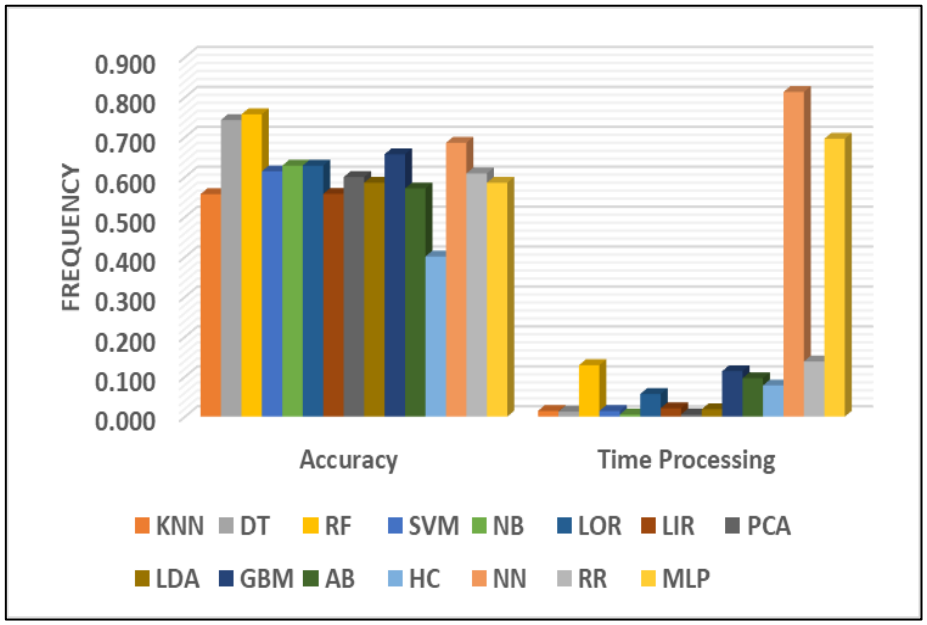
The HD and DSS are two datasets in healthcare that were analyzed in Python scripting to verify accuracy and time limit performance. In the HD dataset, the RF algorithm had the highest accuracy at 0.672, but the NB algorithm had the best processing time, at 0.001. In the DSS dataset, the same algorithms had the highest accuracy of 0.757 and the best processing time, 0.005. More details about the HD and DSS accuracy and time processing were recorded, as shown in table 12 and figure 8.

Table 11: Accuracy and time processing (Sec.) for the HD and DSS datasets.

No.	Algorithms	HD Dataset		DSS Dataset	
		Accuracy	**Time Processing**	**Accuracy**	**Time Processing**
1	KNN	0.574	0.014	0.557	0.014
2	DT	0.541	0.012	0.743	0.013
3	RF	**0.672**	0.012	**0.757**	0.129
4	SVM	0.623	0.014	0.614	0.014
5	NB	0.492	**0.001**	0.629	**0.005**
6	LoR	0.656	0.055	0.629	0.057
7	LiR	0.413	0.021	0.558	0.021
8	MLP	0.541	0.958	0.586	0.696
9	RR	0.554	0.007	0.609	0.138
10	GBM	0.623	0.124	0.657	0.114
11	AB	0.492	0.854	0.571	0.096
12	HC	0.232	0.070	0.401	0.078
13	NN	0.541	0.837	0.686	0.813
14	LDA	0.623	0.019	0.586	0.018
15	PCA	0.573	0.005	0.600	0.006



A- Accuracy



B- Time Processing

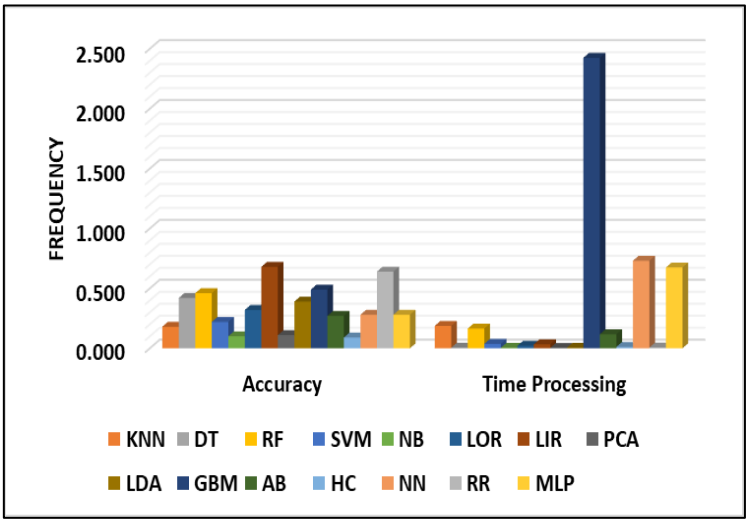
Figure 8: Accuracy and time processing (Sec.) for the HD dataset and DSS dataset in Python.

4.2.2. Educational Datasets Results

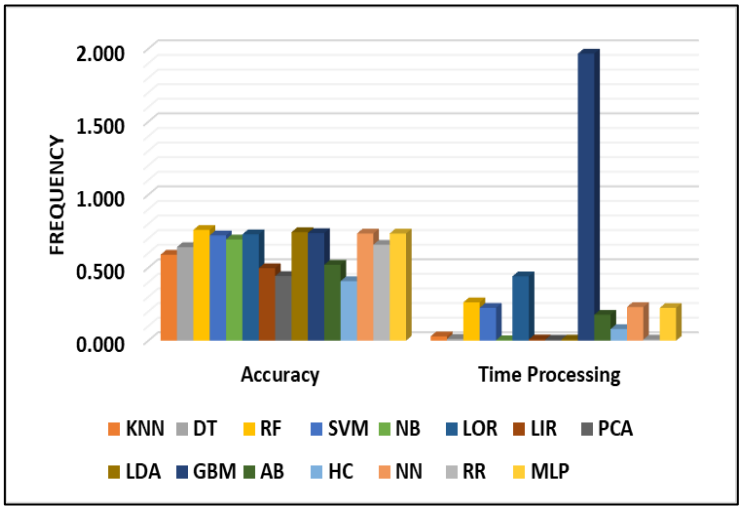
A pair of educational datasets, SP and SWU, were used to verify the Python results' accuracy and time limit. In the SP dataset, the RF algorithm had the highest accuracy at 0.759 but the NB algorithm had the best time processing of 0.003. Also, in the DSS dataset, the LiR algorithm had the highest accuracy of 0.680 but PCA had the best time processing at 0.003. More details about SP and SWU accuracy and time processing were recorded, as shown in table 13 and figure 9.

Table 12: Accuracy and time processing (Sec.) for SP and SWU dataset.

No.	Algorithms	SP Dataset		SWU Dataset	
		Accuracy	Time Processing	Accuracy	Time Processing
1	KNN	0.589	0.029	0.180	0.187
2	DT	0.641	0.012	0.420	0.005
3	RF	0.759	0.263	0.460	0.164
4	SVM	0.721	0.225	0.220	0.037
5	NB	0.694	0.003	0.100	0.004
6	LoR	0.729	0.441	0.320	0.020
7	LiR	0.497	0.009	0.680	0.033
8	MLP	0.734	0.224	0.280	0.674
9	RR	0.657	0.008	0.640	0.006
10	GBM	0.737	1.967	0.490	2.424
11	AB	0.520	0.177	0.270	0.117
12	HC	0.408	0.080	0.090	0.011
13	NN	0.734	0.231	0.280	0.731
14	LDA	0.743	0.008	0.390	0.004
15	PCA	0.442	0.006	0.110	0.003



(A)



(B)

Figure 9: (A) Accuracy and (B) time processing (Sec.) for the SP dataset and SWU dataset in Python.

5. Discussion

This study provides a comparative analysis of a ChatGPT-assisted simulation and traditional Python-based machine learning execution. Unlike prior research that focuses mainly on maximizing accuracy, this work highlights differences in outcomes and computational efficiency between the two approaches, showing ChatGPT's potential as a supportive tool while noting its current limitations in reliability.

The evaluation covered a broad set of algorithms, including clustering (KNN, HC, NN), tree-based (DT, RF, GBM, AB), statistical classifiers (SVM, LoR, NB, RR, LDA), regression and dimensionality reduction (LiR, PCA), and MLP for nonlinear patterns. By contrasting ChatGPT outputs with Python-derived results across these methods, the study demonstrates both the utility and constraints of integrating generative AI in machine learning workflows, consistent with recent findings on the role of AI tools in data analysis [70].

5.1. Heart Disease Dataset

In the ChatGPT tool analysis, the accuracy of the NB, LiR, PCA, AB, NN and MLP algorithms were better, while the time processing of DT, LiR, LDA and AB algorithms were better. In the Python results, the accuracy of the KNN, DT, RF, SVM, LoR, LDA, GBM and RR algorithms were better, while the time processing of the KNN, RF, SVM, NB, LoR, PCA, GBM, NN, RR and MLP algorithms were better. The HC algorithm could not be compared due to the lack of available values in the ChatGPT analysis, as shown in table 14.

Table 13: Selecting best result of accuracy and time processing (Sec.) according experimental types.

No.	Algorithm	ChatGPT Result		Real Code Result	
		Accuracy AVG	Time Processing AVG	Accuracy	Time Processing
1	KNN	0.566	0.035	0.574	0.014
2	DT	0.501	0.006	0.541	0.012
3	RF	0.608	0.274	0.672	0.012
4	SVM	0.581	0.044	0.623	0.014
5	NB	0.554	0.004	0.492	0.001
6	LoR	0.608	0.223	0.656	0.055
7	LiR	0.500	0.008	0.413	0.021
8	MLP	0.559	7.814	0.541	0.958
9	RR	0.538	0.007	0.554	0.007
10	GBM	0.589	1.019	0.623	0.124
11	AB	0.574	0.127	0.492	0.854
12	HC	N/A	N/A	0.232	0.070
13	NN	0.560	2.676	0.541	0.837
14	LDA	0.594	0.007	0.623	0.019
15	PCA	0.576	0.153	0.573	0.005

5.2. Disease Sign and Symptom Dataset

M and MLP algorithms were better, while the time processing of the KNN, DT, SVM, NB, LoR, LiR, LDA, GBM and RR algorithms were better. For the Python application results, the accuracy of all algorithms was better except for GBM and MLP, while the time processing of the RF, PCA, AB, HC and NN algorithms were better, as shown in table 15.

Table 14: Selecting best result of accuracy and time processing (Sec.) according experimental types.

No.	Algorithm	ChatGPT Result		Python Code Result	
		Accuracy AVG	Time Processing AVG	Accuracy	Time Processing
1	KNN	0.552	0.008	0.557	0.014
2	DT	0.738	0.005	0.743	0.013
3	RF	0.753	0.196	0.757	0.129
4	SVM	0.584	0.013	0.614	0.014
5	NB	0.528	0.004	0.629	0.005
6	LoR	0.558	0.038	0.629	0.057
7	LiR	0.549	0.008	0.558	0.021
8	MLP	0.629	6.818	0.586	0.696
9	RR	0.540	0.014	0.609	0.138
10	GBM	0.712	0.105	0.657	0.114
11	AB	0.488	0.110	0.571	0.096
12	HC	0.378	0.366	0.401	0.078
13	NN	0.626	6.665	0.686	0.813
14	LDA	0.558	0.004	0.586	0.018
15	PCA	0.560	0.012	0.600	0.006

5.3. Student Performance Dataset

The accuracy of the DT algorithm was higher, at 0.655, with a processing time of 0.012 in ChatGPT. In comparison, the KNN algorithm demonstrated an accuracy of 0.589, with a processing time of 0.029 in the Python application. Further details regarding the accuracy and processing time of the SP are provided in table 16.

Table 15: Selecting best result of accuracy and time processing (Sec.) according experimental types.

No.	Algorithm	ChatGPT Result		Python Code Result	
		Accuracy AVG	Time Processing AVG	Accuracy	Time Processing
1	KNN	0.588	0.038	0.589	0.029
2	DT	0.655	0.012	0.641	0.012
3	RF	0.757	0.522	0.759	0.263
4	SVM	0.729	0.237	0.721	0.225
5	NB	0.698	0.004	0.694	0.003
6	LoR	0.748	0.360	0.729	0.441
7	LiR	0.470	0.047	0.497	0.009
8	MLP	0.722	11.073	0.734	0.224
9	RR	0.470	0.018	0.657	0.008
10	GBM	0.758	2.102	0.737	1.967
11	AB	0.698	0.252	0.520	0.177
12	HC	0.087	0.069	0.408	0.080
13	NN	0.733	8.293	0.734	0.231
14	LDA	0.758	0.024	0.743	0.008
15	PCA	0.490	0.079	0.442	0.006

5.4. Student World University Dataset

The accuracy of the KNN algorithm was higher, with a value of 0.249 in ChatGPT, whereas the processing time for KNN was more efficient, recorded at 0.187 in the Python application result. In contrast, the DT algorithm demonstrated a higher accuracy of 0.420, with a significantly lower processing time of 0.005 in the Python application (Table 17).

Table 16: Selecting best result of accuracy and time processing (Sec.) according experimental types.

No.	Algorithm	ChatGPT Result		Python Code Result	
		Accuracy AVG	Time Processing AVG	Accuracy	Time Processing
1	KNN	0.249	0.254	0.180	0.187
2	DT	0.409	0.015	0.420	0.005
3	RF	0.469	0.300	0.460	0.164
4	SVM	0.281	0.053	0.220	0.037
5	NB	0.130	0.006	0.100	0.004
6	Lor	0.339	1.027	0.320	0.020
7	Lir	0.669	0.011	0.680	0.033
8	MLP	0.323	21.400	0.280	0.674
9	RR	0.642	0.010	0.640	0.006
10	GBM	0.497	3.292	0.490	2.424
11	AB	0.275	0.154	0.270	0.117
12	HC	0.065	0.029	0.090	0.011
13	NN	0.300	25.477	0.280	0.731
14	LDA	0.396	0.043	0.390	0.004
15	PCA	0.290	0.191	0.110	0.003

Recent studies have demonstrated that ensemble learning methods, particularly RF and GBM, consistently achieve strong predictive performance in healthcare and educational analytics. For instance, Fu [71] compared LR and RF for house price prediction and reported that RF significantly outperformed LR in accuracy, underscoring the superiority of ensemble approaches over linear baselines. Similarly, Li [72] examined RF and XGBoost (XGB) and found that XGB achieved a lower mean absolute error (MAE) than RF, although at the expense of higher computational cost. In the health domain, Adetunji *et al.* [73] applied RF to heart disease prediction and showed that the method effectively handled heterogeneous clinical data, achieving accuracies above 0.90 when cross-validation was employed in the educational field. More recently, Suaza-Medina *et al.* [74] developed a machine learning framework supported by Shapley additive explanations for predicting standardized test outcomes in lagging regions, reporting accuracies above 0.85 while emphasizing model interpretability.

In comparison, the present study evaluated 15 algorithms across health and educational datasets using two distinct pipelines: a Python code-based implementation and a ChatGPT-assisted modeling approach. While previous researches [71-74] primarily emphasized maximizing predictive accuracy through optimized ensemble and hybrid methods, our findings provide a broader perspective by jointly considering predictive accuracy and computational efficiency. For example, in the health-related datasets, RF achieved 0.672 accuracy on the HD dataset and 0.757 on the DSS dataset, whereas NB offered the fastest inference time (≈0.001–0.005 seconds). In educational datasets, RF reached 0.759 accuracy on SP, while regression methods yielded higher performance on SWU, with LiR achieving up to 0.862. These results align with the prior evidence indicating that ensemble methods are generally strong performers [71-74], yet the unique contribution of this work lies in demonstrating the trade-offs between model accuracy, execution time, and evaluation protocols under different computational settings. By integrating both conventional code-driven and AI-assisted pipelines, this study advances the discussion beyond accuracy benchmarks alone and highlights practical considerations for deploying machine learning models in real-world health and education applications.

This work faces several important limitations. The evaluation of response speed and scalability across larger datasets is still insufficient, as current experiments do not fully capture real-world demands when comparing AI-based data processing tools with ChatGPT.

Another limitation of this study is the inability to directly differentiate and compare our results with other works that used ChatGPT-4o on similar datasets and models. A fair comparison is

challenging because the same datasets are often evaluated under different conditions, using various AI chatbots or alternative AI tools. In particular, performance metrics such as accuracy may vary depending on temporal factors (e.g., different testing periods) or contextual factors (e.g., geographic or system-specific settings). These variations make it difficult to find a consistent baseline for evaluation, thereby limiting the strength of the comparative analysis.

In addition, ChatGPT shows vulnerability in authentication during dataset generation; this issue could be better understood and managed within a multilevel security identification framework. These challenges highlight the need for future research to strengthen ChatGPT's reliability and adaptability, while also guiding users to remain cautious about its current constraints.

6. Conclusions

This study investigates the error rate and performance variability of ChatGPT's responses across different time zones. It identifies specific periods when certain ML algorithms fail to produce results to determine the most effective time for executing each algorithm through repeated trials. This study emphasizes the need for the accurate evaluation of ChatGPT-generated results by comparing them with traditional performance metrics. It outlines the main objectives and introduces an AI framework for identifying optimal analysis times, ultimately helping AI tool users achieve more accurate and reliable results.

In conclusion, the findings revealed that the proposed time-complexity approach successfully balanced accuracy and processing time. Extensive experiments confirmed its effectiveness in identifying the optimal trade-off between these factors when comparing Python-based implementations with ChatGPT analysis across 15 classifiers, including KNN, DT, RF, SVM, and NB. This study contributes to improving service efficiency and time management in ChatGPT applications. Additionally, some ML algorithms occasionally encounter issues and fail to provide results in ChatGPT, such as PC and HC. Thus, this analysis showed that AI tools like chatbots can have unavailable responses at certain times during analysis. It also revealed issues such as generating incorrect information and repeating outputs, highlighting the need for better timing and reliability in their performance.

Future research will aim to enhance response times and assess performance on larger datasets, focusing on a comparison between AI-driven data processing tools and ChatGPT. Efforts will be directed towards applying parameter tuning to enhance classification algorithms for new, complex datasets in the healthcare and education sectors. This will allow for a more thorough evaluation of the limitations inherent in ChatGPT, particularly in relation to detecting classification errors. Furthermore, ChatGPT has exhibited authentication challenges during dataset generation, which could be analyzed within a multi-level security framework. To address these issues, multi-objective metaheuristic algorithms will be utilized to improve both performance and accuracy, ultimately advancing the capabilities of ChatGPT. Consequently, future studies must prioritize identifying the limitations of these AI tools and addressing key areas for improvement.

Author contributions: Bnar Kamaran Arif: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, **Aso M. Aladdin:** Conceptualization, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Data availability: Data will be available upon reasonable request by Author.

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding: The authors did not receive support from any organization for the conducting of the study.

References

[1] S. Zhu *et al.*, "Intelligent computing: The latest advances, challenges, and future," *Intelligent Computing*, vol. 2, pp. 1–88, Jan. 2023, doi: 10.34133/icomputing.0006.

[2] Y. K. Dwivedi, A. Sharma, N. P. Rana, M. Giannakis, P. Goel, and V. Dutot, "Evolution of artificial intelligence research in Technological Forecasting and Social Change: Research topics, trends, and future directions," *Technological Forecasting and Social Change*, vol. 192, p. 122579, Jul. 2023, doi: 10.1016/j.techfore.2023.122579.

[3] T. R. G. Babu, U. Saravanakumar, and B. Pattanaik, *Computational Imaging and Analytics in Biomedical Engineering*. New York: Apple Academic Press, 2024. doi: 10.1201/9781032669687.

[4] G. Jeffson Sagala and Y. T. Samuel, "Sentiment analysis on ChatGPT App reviews on google play store using random forest algorithm, support vector machine and naïve bayes," *International Journal of Engineering Business and Social Science*, vol. 2, no. 04, pp. 1194–1204, Mar. 2024, doi: 10.58451/ijebss.v2i04.148.

[5] M. Spring, J. Faulconbridge, and A. Sarwar, "How information technology automates and augments processes: Insights from Artificial-Intelligence-based systems in professional service operations," *Journal of Operations Management*, vol. 68, no. 6–7, pp. 592–618, Sep. 2022, doi: 10.1002/joom.1215.

[6] L. De Angelis *et al.*, "ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health," *Frontiers in Public Health*, vol. 11, Apr. 2023, doi: 10.3389/fpubh.2023.1166120.

[7] OpenAI *et al.*, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774.

[8] B. Zohur and F. M. Rahmani, "ChatGPT vs Chatbots unleashing the power of conversational AI," *Journal of Material Sciences & Manufacturing Research*, vol. 4, no. 5, pp. 1–5, 2023, doi: 10.47363/JMSMR/2023(4)158.

[9] N. Khan, Z. Khan, A. Koubaa, M. K. Khan, and R. bin Salleh, "Global insights and the impact of generative AI-ChatGPT on multidisciplinary: a systematic review and bibliometric analysis," *Connection Science*, vol. 36, no. *1, Article 2353630*, Dec. 2024, doi: 10.1080/09540091.2024.2353630.

[10] A. M. Rizki, B. Bustami, and S. F. Anshari, "Comparison of support vector machine and NaÃ¯ve Bayes algorithms in sentiment analysis of Tiktokshop application user reviews," *Journal of Renewable Energy, Electrical, and Computer Engineering*, vol. 5, no. 1, pp. 18–29, Mar. 2025, doi: 10.29103/jreece.v5i1.21342.

[11] N. Ansari, V. Babaei, and M. M. Najafpour, "Enhancing catalysis studies with chat generative pre-trained transformer (ChatGPT): Conversation with ChatGPT," *Dalton Transactions*, vol. 53, no. 8, pp. 3534–3547, 2024, doi: 10.1039/D3DT04178F.

[12] A. S. Pamungkas and N. Cahyono, "Analisis Sentimen Review ChatGPT di Play Store menggunakan Support Vector Machine dan K-Nearest Neighbor," *Edumatic: Jurnal Pendidikan Informatika*, vol. 8, no. 1, pp. 1–10, Jun. 2024, doi: 10.29408/edumatic.v8i1.24114.

[13] Y. Akbar, A. N. H. Regita, Sugiyono, and T. Wahyudi, "Analisa Sentimen Pada Media Sosial'X' Pencarian keyword ChatGPT menggunakan algoritma K-Nearest Neighbors (KNN)," *Jurnal Indonesia : Manajemen Informatika dan Komunikasi*, vol. 5, no. 3, pp. 3291–3305, Sep. 2024, doi: 10.35870/jimik.v5i3.1016.

[14] N. Cong-Lem, A. Soyoof, and D. Tsering, "A Systematic review of the limitations and associated opportunities of ChatGPT," *International Journal of Human–Computer Interaction*, vol. 41, no. 7, pp. 3851–3866, May 2024, doi: 10.1080/10447318.2024.2344142.

[15] A. A. Alqudah *et al.*, "Evaluating accuracy and reproducibility of ChatGPT responses to patient-based questions in Ophthalmology: An observational study," *Medicine*, vol. 103, no. 32, p. e39120, Aug. 2024, doi: 10.1097/MD.0000000000039120.

[16] A. Keles, O. G. Illeez, B. Erbagci, and E. Giray, "Artificial intelligence-generated responses to frequently asked questions on coccydynia: Evaluating the accuracy and consistency of GPT-4o's performance," *Archives of Rheumatology*, vol. 40, no. 1, pp. 63–71, Mar. 2025, doi: 10.46497/ArchRheumatol.2025.10966.

[17] J. M. Lapates, M. D. G. Dacer, and D. N. Gaylo, "Analysis of student output on the use of ChatGPT: A predictive model approach," *International Journal of Engineering Trends and Technology*, vol. 72, no. 10, pp. 216–224, Oct. 2024, doi: 10.14445/22315381/IJETT-V72I10P121.

[18] D. O. Hasan, A. M. Aladdin, H. S. Talabani, T. A. Rashid, and S. Mirjalili, "The fifteen puzzle—a new approach through hybridizing three heuristics methods," *Computers*, vol. 12, no. 1, p. 11, Jan. 2023, doi: 10.3390/computers12010011.

[19] M. Maktabdar Oghaz, L. Babu Saheer, K. Dhame, and G. Singaram, "Detection and classification of ChatGPT-generated content using deep transformer models," *Frontiers in Artificial Intelligence*, vol. 8, Article 1458707, Apr. 2025, doi: 10.3389/frai.2025.1458707.

[20] B. Zhao, W. Jin, J. Del Ser, and G. Yang, "ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification," *Neurocomputing*, vol. 557, p. 126708, Nov. 2023, doi: 10.1016/j.neucom.2023.126708.

[21] I. Tri Julianto, D. Kurniadi, Y. Septiana, and A. Sutedi, "Alternative text pre-processing using Chat GPT Open AI," *Jurnal Nasional Pendidikan Teknik Informatika*, vol. 12, no. 1, pp. 67–77, Mar. 2023, doi: 10.23887/janapati.v12i1.59746.

[22] Y. Guo and C. Wang, "Improvement path of legal system related to chatgpt application combined with decision tree algorithm," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, pp. 1–18, Jan. 2024, doi: 10.2478/amns-2024-1396.

[23] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, vol. 2, pp. 121–154, 2023, doi: 10.1016/j.iotcps.2023.04.003.

[24] A. Haleem, M. Javaid, and R. P. Singh, "Exploring the competence of ChatGPT for customer and patient service management," *Intelligent Pharmacy*, vol. 2, no. 3, pp. 392–414, Jun. 2024, doi: 10.1016/j.ipha.2024.03.002.

[25] W. B. L. G. A. A. and S. L. A. Koubaa, " Exploring ChatGPT capabilities and limitations: A survey," *IEEE Access*, vol. 11, pp. 118698–118721, 2023, doi: 10.1109/ACCESS.2023.3326474.

[26] T. H. Kung *et al.*, "Performance of ChatGPT on USMLE: potential for ai-assisted medical education using large language models," *PLOS Digital Health*, vol. 2, no. 2, p. e0000198, Feb. 2023, doi: 10.1371/journal.pdig.0000198.

[27] A. Azaria, R. Azoulay, and S. Reches, "ChatGPT is a remarkable tool—for experts," *Data Intelligence*, vol. 6, no. 1, pp. 240–296, Feb. 2024, doi: 10.1162/dint_a_00235.

[28] Y. K. Dwivedi *et al.*, "Opinion Paper: 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102642, Aug. 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[29] A. M. Aladdin and T. A. Rashid, "LEO: Lagrange elementary optimization," *Neural Computing Application*, May 2025, doi: 10.1007/s00521-025-11225-2.

[30] G. P. B. P. V. Prathamesh Muzumdar1, "An empirical comparison of machine learning models for student's mental health illness assessment," *Asian Journal of Computer and Information Systems*, vol. 10, no. 1, Feb. 2022.

[31] C. K. Subasri and V. Jeyakumar, "Comparative analysis of machine learning algorithms for diabetes prediction using Real-time data-set," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, Jul. 2023, pp. 1–5. doi: 10.1109/ICCCNT56998.2023.10306851.

[32] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, Mar. 2023, doi: 10.3390/healthcare11060887.

[33] D. R. Nugroho, N. Idris, K. Kurniawan, D. Fitrianah, E. Irwansyah, and G. P. Kusuma, "Logistic regression and random forest comparison in predicting students' qualification based on students' half-semester performance," in *2023 11th International Conference on Information and Communication Technology (ICoICT)*, IEEE, Aug. 2023, pp. 214–219. doi: 10.1109/ICoICT58202.2023.10262783.

[34] N. Hasan, J. A. Polin, Md. R. Ahmmed, Md. M. Sakib, Md. F. Jahin, and Md. M. Rahman, "A novel approach to analyzing the impact of AI, ChatGPT, and chatbot on education using machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 4, pp. 2951–2958, Aug. 2024, doi: 10.11591/eei.v13i4.7158.

[35] H. Chen *et al.*, "Multi role ChatGPT framework for transforming medical data analysis," *Scientific Reports*, vol. 14, no. 1, p. 13930, Jun. 2024, doi: 10.1038/s41598-024-64585-5.

[36] B. Basheer and M. Obaidi, "Enhancing healthcare data classification: Leveraging machine learning on ChatGPT-generated datasets," *International Journal of Advances in Applied Computational Intelligence*, vol. 5, no. 2, pp. 34–45, 2024, doi: 10.54216/IJAACI.050203.

[37] A. Rejeb, K. Rejeb, A. Appolloni, H. Treiblmaier, and M. Iranmanesh, "Exploring the impact of ChatGPT on education: A web mining and machine learning approach," *The International Journal of Management Education*, vol. 22, no. 1, p. 100932, Mar. 2024, doi: 10.1016/j.ijme.2024.100932.

[38] G. Narasimhan and A. Victor, "Empirical analysis of predicting heart disease using diverse datasets and classification procedures of machine learning," *Ain Shams Engineering Journal*, vol. 16, no. 8, p. 103470, Aug. 2025, doi: 10.1016/j.asej.2025.103470.

[39] S. Malik *et al.*, "Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution," *Scientific Reports*, vol. 15, no. 1, p. 8738, Mar. 2025, doi: 10.1038/s41598-025-92324-x.

[40] U. J. Nzenwata, E. Edwin, E. A. Chukwu, D. Osilaja, J. O. Hinmikaiye, and C. Enyinnah, "Extra trees model for heart disease prediction," *Journal of Data Analysis and Information Processing*, vol. 13, no. 02, pp. 125–139, 2025, doi: 10.4236/jdaip.2025.132008.

[41] K. Zhou *et al.*, "A machine learning model for predicting acute respiratory distress syndrome risk in patients with sepsis using circulating immune cell parameters: a retrospective study," *BMC Infectious Diseases*, vol. 25, no. 1, p. 568, Apr. 2025, doi: 10.1186/s12879-025-10974-8.

[42] J. A. Benítez-Andrades, C. Prada-García, N. Ordás-Reyes, M. E. Blanco, A. Merayo, and A. Serrano-García, "Enhanced prediction of spine surgery outcomes using advanced machine learning techniques and oversampling methods," *Health Information Science and Systems*, vol. 13, no. 1, p. 24, Mar. 2025, doi: 10.1007/s13755-025-00343-9.

[43] A. A. H. Amin, A. M. Aladdin, D. O. Hasan, S. R. Mohammed-Taha, and T. A. Rashid, "Enhancing algorithm selection through comprehensive performance evaluation: statistical analysis of stochastic algorithms," *Computation*, vol. 11, no. 11, p. 231, Nov. 2023, doi: 10.3390/computation11110231.

[44] P. B. Serafim, P. Crescenzi, G. Gezici, E. Cappuccio, S. Rinzivillo, and F. Giannotti, "Exploring large language models capabilities to explain decision trees," vol. 386, pp. 64–72, 2024. doi: 10.3233/FAIA240183.

[45] Hemin Sardar Abdulla, Azad A. Ameen, Sarwar Ibrahim Saeed, and Tarik A. Rashid, "MRSO: Balancing exploration and exploitation through," *Algorithms*, vol. 17, no. 9, 2024, doi: 10.3390/a17090423.

[46] I. Lintang, A. D. Lestari, and B. Prasetiyo, "Application of k-nearest neighbor algorithm in classification of engine performance in car companies using Rapidminer," *Journal of Student Research Exploration*, vol. 2, no. 2, pp. 120–130, Jul. 2024, doi: 10.52465/josre.v2i2.345.

[47] A. Sabir, H. A. Ali, and M. A. Aljabery, "ChatGPT Tweets sentiment analysis using machine learning and data classification," *Informatica*, vol. 48, no. 7, pp. 103–112, May 2024, doi: 10.31449/inf.v48i7.5535.

[48] S. Kodera, O. Yokoi, M. Kaneko, Y. Sato, S. Ito, and K. Hata, "Age estimation from blood test results using a random forest model," Feb. 06, 2024. doi: 10.1101/2024.02.06.24302114.

[49] J. P. Munggaran, A. A. Alhafidz, M. Taqy, D. A. R. Agustini, and M. Munawir, "Sentiment analysis of twitter users' Opinion data regarding the use of chatgpt in education," *Journal of Computer Engineering, Electronics and Information Technology*, vol. 2, no. 2, pp. 75–88, Jun. 2023, doi: 10.17509/coelite.v2i2.59645.

[50] C. Duncan and I. Mcculloh, "Unmasking bias in ChatGPT responses," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, New York, NY, USA: ACM, Nov. 2023, pp. 687–691. doi: 10.1145/3625007.3627484.

[51] S. Rabbani, D. Safitri, F. Try Puspa Siregar, R. Rahmaddeni, and L. Efrizoni, "Evaluation of support vector machine, naive bayes, decision tree, and gradient boosting algorithms for sentiment analysis on ChatGPT twitter dataset," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 7, no. 1, pp. 11–21, Nov. 2023, doi: 10.24014/ijaidm.v7i1.24662.

[52] O. Shobayo, S. Adeyemi-Longe, O. Popoola, and B. Ogunleye, "innovative sentiment analysis and prediction of stock price using finbert, gpt-4 and logistic regression: A data-driven approach," *Big Data and Cognitive Computing*, vol. 8, no. 11, p. 143, Oct. 2024, doi: 10.3390/bdcc8110143.
[53] C. Zhong and J. B. Kim, "Teaching case teaching business students logistic regression in r with the aid of ChatGPT," *Journal of Information Systems Education*, vol. 35, no. 2, pp. 138–143, 2024, doi: 10.62273/DYLI2468.
[54] S.-L. Cheng *et al.*, "Comparisons of quality, correctness, and similarity between chatgpt-generated and human-written abstracts for basic research: Cross-sectional study," *Journal of Medical Internet Research*, vol. 25, p. e51229, Dec. 2023, doi: 10.2196/51229.
[55] C. A. Sequeira and E. M. Borges, "enhancing statistical education in chemistry and STEAM Using JAMOVI. Part 2. Comparing dependent groups and principal component analysis (PCA)," *Journal of Chemical Education*, vol. 101, no. 11, pp. 5040–5049, Nov. 2024, doi: 10.1021/acs.jchemed.4c00342.
[56] S. Sotirov and I. Dimitrov, "Application of machine learning algorithms for prediction of tumor T-cell immunogens," *Applied Sciences*, vol. 14, no. 10, p. 4034, May 2024, doi: 10.3390/app14104034.
[57] A. Li, Y. Wang, and H. Chen, "AI driven cardiovascular risk prediction using NLP and large language models for personalized medicine in athletes," *SLAS Technology*, vol. 32, p. 100286, Jun. 2025, doi: 10.1016/j.slast.2025.100286.
[58] H. Yildiz Durak, F. Eğin, and A. Onan, "A comparison of human-written versus AI -generated text in discussions at educational settings: investigating features for ChatGPT , Gemini and BingAI," *European Journal of Education*, vol. 60, no. 1, Mar. 2025, doi: 10.1111/ejed.70014.
[59] F. Hauth *et al.*, "electronic patient-reported outcome measures in radiation oncology: Initial experience after workflow implementation," *JMIR mHealth and uHealth*, vol. 7, no. 7, p. e12345, Jul. 2019, doi: 10.2196/12345.
[60] K. Abdalgader, A. A. Matroud, and K. Hossin, "Experimental study on short-text clustering using transformer-based semantic similarity measure," *PeerJ Computer Science*, vol. 10, p. e2078, May 2024, doi: 10.7717/peerj-cs.2078.
[61] X. Zhao, "Capitalized comparison of three machine learning models: Linear model, decision tree, neural network," *Highlights in Science, Engineering and Technology*, vol. 85, pp. 790–796, Mar. 2024, doi: 10.54097/pxcpca72.
[62] H. Luo and H. Kong, "ChatGPT plagiarism in the academic field: Exploration and analysis of plagiarism effects," *International Conference on Machine Learning and Intelligent Computing*, vol. 245, pp. 1–11, 2024, Accessed: Aug. 14, 2025. [Online]. Available: https://proceedings.mlr.press/v245/haiqiong24a.html.
[63] H. Fakour and M. Imani, "Socratic wisdom in the age of AI: a comparative study of ChatGPT and human tutors in enhancing critical thinking skills," *Frontiers in Education (Lausanne)*, vol. 10, pp. 1–12, 2025, doi: 10.3389/feduc.2025.1528603.
[64] R. Krimsony, "UCI Heart Disease Data," *Kaggle*. Accessed: Jul. 22, 2025. [Online]. Available: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data.
[65] L. Tharmalingam, "Disease Symptoms and Patient Profile Dataset," *Kaggle*. Accessed: Jul. 22, 2025. [Online]. Available: https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset.
[66] R. Elkharoua, "Students' Performance Dataset," *Kaggle*. Accessed: Jul. 22, 2025. [Online]. Available: https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset.
[67] Larsen0966, "Student Performance Data Set," *Kaggle*. Accessed: Jul. 22, 2025. [Online]. Available: https://www.kaggle.com/datasets/larsen0966/student-performance-data-set.
[68] O. B. Elhalid, Z. Alm Alhelal, and S. HASSAN, "Exploring the fundamentals of python programming: A comprehensive guide for beginners," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4612765.
[69] C. J. Weiss, "Visualizing protein big data using Python and Jupyter notebooks," *Biochemistry and Molecular Biology Education*, vol. 50, no. 5, pp. 431–436, Sep. 2022, doi: 10.1002/bmb.21621.
[70] O. Lederman *et al.*, "Promises and perils of generative artificial intelligence: a narrative review informing its ethical and practical applications in clinical exercise physiology," *BMC Sports Science, Medicine and Rehabilitation*, vol. 17, no. 1, p. 131, May 2025, doi: 10.1186/s13102-025-01182-7.
[71] Y. Fu, "A comparative study of house price prediction using linear regression and random forest models," *Highlights in Science, Engineering and Technology*, vol. 107, pp. 96–103, Aug. 2024, doi: 10.54097/vcy5n584.
[72] H. Li, "House price prediction and analysis based on random forest and XGBoost models," *Highlights in Business, Economics and Management*, vol. 21, pp. 934–938, Dec. 2023, doi: 10.54097/hbem.v21i.14837.
[73] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House price prediction using random forest machine learning technique," *Procedia Computer Science*, vol. 199, pp. 806–813, 2022, doi: 10.1016/j.procs.2022.01.100.
[74] M. Suaza-Medina, R. Peñabaena-Niebles, and M. Jubiz-Diaz, "A model for predicting academic performance on standardised tests for lagging regions based on machine learning and Shapley additive explanations," *Scientific Reports*, vol. 14, no. 1, p. 25306, Oct. 2024, doi: 10.1038/s41598-024-76596-3.