Impact of Lifestyle Influences on Heart Health

Neha Gupta #23265481
neha.gupta3@mail.dcu.ie
Dublin City University

Omkar Khaladkar #23262297
omkar.khaladkar2@mail.dcu.ie
Dublin City University

Vaishnavi Kulkarni #23266282
vaishnavi.kulkarni2@mail.dcu.ie
Dublin City University

Shreya Ketkar #23262829
shreya.ketkar2@mail.dcu.ie
Dublin City University

Stanley Johnson #23260879
stanley.johnson7@mail.dcu.ie
Dublin City University

I. Abstract

Heart disease's prevalence as a leading global killer highlights an urgent need for improved predictive analytics, particularly given the high incidence of modifiable lifestyle risk factors such as obesity and smoking. This study taps into machine learning's potential to predict heart disease more accurately by analyzing clinical and lifestyle-related attributes. Addressing the challenge of a significantly imbalanced dataset—characteristic of the condition's lower prevalence—our methodology re-calibrates the balance between sensitivity and specificity, favoring metrics such as the F1-score and AUC-ROC over raw accuracy. Despite inherent multicollinearity within the lifestyle variables, our robust feature engineering process enabled the effective use of three key machine learning classifiers, including Logistic Regression and Gradient Boosting, yielding models with an improved predictive recall. This research paves the way for targeted interventions and underscores the transformative impact of a data-centric approach in combating heart disease.

II. Introduction

The prevalence of heart disease as a leading cause of mortality across diverse populations in the United States underscores a critical public health issue that demands immediate attention [1]. Notably, this condition does not discriminate, affecting individuals across a spectrum of racial and ethnic groups, with a particularly pronounced impact on African Americans, American Indians and Alaska Natives, and whites. In the wake of the global COVID-19 pandemic, the resultant societal shifts have only magnified the importance of understanding and addressing the myriad factors contributing to heart disease [8].

Central to the discourse on mitigating heart disease is the recognition of its multifaceted risk factors, categorized broadly into those that are non-modifiable, such as age, gender, and genetics, and those that are modifiable, including lifestyle choices and certain health conditions. Specifically, hypertension, high cholesterol, and smoking are identified as three key modifiable risk factors, affecting approximately 47% of the American populace [1]. The interplay of these risk factors, alongside diabetes, obesity, and inadequate physical activity, presents a complex web of challenges that significantly elevate the risk for heart disease.

Moreover, lifestyle choices such as poor diet, lack of physical activity, excessive alcohol consumption, and tobacco use have been directly linked to an increased risk of heart disease. These behaviors exacerbate the effects of the aforementioned health conditions and, when coupled with genetic predispositions and environmental influences, create a potent recipe for heart disease.

The burgeoning field of data science offers a promising avenue for dissecting and understanding the impact of lifestyle choices on heart disease. Through the application of machine learning and predictive analytics, researchers are now equipped to sift through vast datasets to identify patterns and correlations that can inform more effective prevention and treatment strategies.

This paper seeks to explore the nexus between lifestyle factors and heart disease, leveraging state-of-the-art data science methodologies to unearth insights that could pave the way for personalized and proactive interventions. By doing so, it aims to contribute to the broader goal of reducing the

[1] https://www.cdc.gov/heartdisease/

prevalence of heart disease and enhancing the quality of life for those affected.

This paper is organized as follows: Section III presents a comprehensive Literature Review, discussing relevant studies and prior work in the domain of heart health. Section IV, Methodology, is divided into three primary subsections: (i) Dataset Summary, which outlines the CDC Heart Disease dataset (ii) Exploratory Data Analysis, detailing our initial findings and observations,and (iii) Data Preprocessing and Feature Engineering, describing the techniques applied to prepare the data for modeling (iv) CRISP-DM Methodology, elucidates our analytical process guided by the Cross-Industry Standard Process for Data Mining (v) Modeling Approach, provides an overview of the predictive models and techniques. Section V details the Evaluation Methods used to assess the effectiveness of our predictive models. Finally, Section VI offers Conclusions drawn from this study and outlines potential avenues for Future Research.

III. LITERATURE REVIEW

Cardiovascular health is a critical aspect of overall well-being, and various lifestyle factors play a significant role in determining heart health outcomes. The study by [4] explores how unhealthy lifestyle components impact cardiorespiratory fitness and heart rate recovery among medical students. This research highlights the essential roles of physical activity and maintaining a healthy weight in cardiovascular health, setting the stage for promoting healthier behaviors early in medical careers to mitigate cardiometabolic risks.

Transitioning from individual lifestyle modifications to broader epidemiological [2] insights, extensive studies reviewed by sources such as the BMJ[3] and the CDC[4] emphasize traditional risk factors like hypertension and high cholesterol. Moreover, they point to lifestyle habits, including smoking and diet, as pivotal elements in heart disease prevalence. Complementing this traditional view, [10] provide a decade-long study from Japan that evaluates the predictive power of these factors on cardiovascular disease (CVD). They propose that integrating lifestyle factors into prediction models enhances early detection capabilities, a crucial step towards preemptive interventions.

However, the application of machine learning techniques to predict heart disease presents both opportunities and challenges, particularly regarding data quality. [7] and [12] discuss how class imbalance in health-related data can skew the performance of machine learning models. They note the effectiveness of approaches like oversampling in managing these imbalances, vital for ensuring robust model training and accurate predictions.

Furthermore, the detailed analysis by [12] explores the impact of class imbalance correction techniques on clinical prediction models, assessing 1,566 models from observational health data. Their findings suggest that while techniques such as random oversampling and undersampling do not significantly enhance the AUROC, the use of more sophisticated methods like SMOTE could potentially yield better outcomes.

Recent advancements underscore the increasing utility of machine learning in enhancing heart disease prediction. [2] review six ML algorithms, with logistic regression and AdaBoost Classifier emerging as particularly effective, especially when combined with GridsearchCV for hyperparameter optimization. These techniques not only improve the accuracy of predictions but also highlight the necessity for continuous methodological innovations to refine disease classification systems.

In conclusion, the integration of lifestyle factors with advanced machine learning models presents a promising frontier for enhancing the predictability and prevention of heart disease. This review emphasizes the need for ongoing innovation in data handling and model optimization to better harness the potential of these technologies in public health contexts.

IV. METHODOLOGY

A. ***Dataset Summary for CDC Heart Disease***

1) Overview: The CDC Heart Disease Dataset[5] is a comprehensive collection of health-related variables aimed at understanding the risk factors associated with heart disease. It consists of 319,795 individual records and includes 18 distinct columns, encompassing both numerical and categorical data types.

The dataset covers a wide range of factors, including demographic details (e.g., Age, Gender, Race), lifestyle choices (e.g., Smoking, Alcohol Drinking, Physical Activity), medical history (e.g., Stroke, Asthma, Kidney Disease, Skin Cancer), and health status indicators (e.g., BMI, PhysicalHealth, MentalHealth, SleepTime).

Numerical Features: Four variables are numerical, representing key health metrics such as BMI, PhysicalHealth, MentalHealth, and SleepTime, which offer quantifiable insights into individual health profiles.

Categorical Features: Fourteen variables are categorical, including HeartDisease (the target variable), with categories

[2] https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/1-what-epidemiology

[3] https://www.bmj.com/

[4] https://www.cdc.gov/nchs/data/hus/2017/019.pdf

[5] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

detailing health conditions, behaviors, and demographic characteristics.

2) Data Acquisition: The dataset was sourced from the Centers for Disease Control and Prevention's database. The data reflects information collected from a broad demographic to gauge the prevalence of heart disease and associated risk factors.

3) Provenance and Collection Methods: The data was compiled through surveys and health evaluations conducted by the CDC, using standardized questionnaires and health assessment protocols to ensure consistency and reliability.

B. Exploratory Data Analysis

The dataset is from the year 2020, encompassing 319,795 records across 18 variables, involved both univariate and bivariate analyses, focused on identifying key trends and relationships pertinent to heart disease.

Univariate analysis underscored the distributions of four numerical variables. BMI was approximately normally distributed, suggesting a relatively healthy weight status among the population. Physical Health and Mental Health were both right-skewed, indicating a majority of the population reported minimal health issues. Sleep Time displayed a near-normal distribution with a slight right skew, denoting most individuals had consistent sleep patterns, with fewer reports of excessive sleep.

The bivariate analysis emphasized relationships between categorical variables and heart disease prevalence. Smoking, alcohol consumption, stroke history, and difficulty walking were closely observed (TABLE I). The analysis revealed that 36.2% of individuals smoked, and smokers had a slightly higher prevalence of heart disease. Alcohol consumption was low, with only 6.5% affirming consumption, while stroke survivors showed a higher prevalence of heart disease. Difficulty walking was reported by 10.8% of individuals, correlating with an elevated heart disease rate.

Gender distribution was near-equal, yet physical activity levels varied, with 72.4% reporting regular activity, and this was inversely related to heart disease occurrence. General health perceptions ranged widely, with ”Very Good” and ”Excellent” accounting for a combined 54.5% of responses, indicative of a general positive self-assessed health status.

Age-wise analysis disclosed a gradual increase in heart disease prevalence with age, with those 50 or older (TABLE II). Racial demographics indicated a higher prevalence of heart disease among White individuals, at 7.4% , which needs to be contextualized within the racial composition of the dataset.

Category	Sub-Category	Heart Disease -Yes (%)	Heart Disease - No (%)
Smoking	Yes	36.2	5.0
	No	55.2	3.5
Alcohol Drinking	Yes	0.4	6.5
	No	8.7	84.2
Stroke	Yes	1.5	2.5
	No	7.6	89
DiffWalking	Yes	3.3	11.4
	No	5.7	80.7
Sex	female	3.7	49.2
	Male	5.3	42.5
Physical Activity	Yes	5.8	70.6
	No	3.3	20.4
Generalhealth	Very Good	1.8	33.9
	fair	2.3	8.6
	Good	3.2	27.1
	Poor	1.3	2.4
	Excellent	0.5	20.4
Asthma	Yes	1.6	12.5
	No	7.4	79.6
KidneyDisease	Yes	1.1	2.6
	No	7.5	88.8
Skin Cancer	Yes	1.6	8.1
	No	7.4	83.7

Table I
PREVALENCE OF HEART DISEASE BY DEMOGRAPHIC AND LIFESTYLE CHOICES

Age Category	Heart Disease: Yes (%)	Heart Disease: No (%)
18-24	0	6.6
25-29	0	5.4
30-34	0.1	5.9
35-39	0.1	6.4
40-44	0.2	6.4
45-49	0.2	6.6
50-54	0.5	7.4
55-59	0.7	8.4
60-64	1.1	9.2
65-69	1.4	9.1
70-74	1.6	8.1
75-79	1.3	5.5
80 or older	1.8	5.9

Table II
PREVALENCE OF HEART DISEASE BY AGE CATEGORY

C. Data Pre-processing and Feature Engineering

Data preprocessing encompassed several critical steps to ensure the data quality and suitability for our model:

A. Encoding Categorical Variables: The 14 categorical variables underwent encoding to translate into a format amenable for modeling. Methods such as auto label encoding was applied based on the categorical variable's intrinsic nature.

B. Handling Missing Values: Upon thorough examination of the dataset, which consisted of 3 million records, only a minimal fraction—18,000 entries—exhibited missing values. In light of the relatively negligible amount, careful imputation strategies were implemented to preserve the dataset's integrity without necessitating the exclusion of these records.

C. Addressing Data Imbalance: The target variable, 'Heart Disease,' was highly skewed, with a disproportionate distribution of 91% 'No' cases to 9% 'Yes' cases. To counter this imbalance and foster robust model training, we explored three distinct balancing strategies: undersampling the majority class, oversampling the minority class, and employing SMOTE (Synthetic Minority Over-sampling Technique). Each method was evaluated to determine its impact on model performance, with SMOTE generating synthetic examples of the underrepresented class to achieve parity in class distribution. These techniques were meticulously applied after partitioning the data to avoid introducing synthetic data into the test set, which could bias our model's performance metrics. Such rigorous balancing is indispensable to ensure a more equitable and accurate learning process.

D. Identifying multicollinearity: In preparation for feature selection within our classification models, we conducted a Variance Inflation Factor (VIF) [6] analysis to ascertain the presence of multicollinearity among predictors. Results indicated minimal multicollinearity, with most variables exhibiting VIF scores around 1. Nevertheless, a subset of features—specifically Age Category, Physical Activity, BMI, Sleep Time, and Race—showed VIF scores exceeding 1, signifying potential multicollinearity concerns.

To effectively manage this issue, particularly given the mix of continuous (like BMI) and discrete variables, we employed Principal Component Analysis (PCA) [5] alongside Factor Analysis of Mixed Data (FAMD) [3]. These dimensionality reduction techniques are well-suited to handle datasets with heterogeneous variable types. PCA was utilized to distill the continuous variables, while FAMD was applied to the entire feature set, harmonizing the continuous and discrete predictors into principal components. This approach not only mitigated the influence of multicollinearity but also streamlined the dataset.

E. Dataset Splitting: The processed dataset was then split into training and test sets to validate the performance of the predictive models. In this study, the dataset was divided with an 80/20 split—80% was allocated for training (and within it, for applying sampling methods), and the remaining 20% is test dataset.

F. Outlier Detection: In our analysis, we identified outliers in several key features such as BMI, Mental Health, Sleep Time, and Physical Health. Upon consulting authoritative studies, we concluded that these outliers represent realistic variations:

BMI: Variability across populations and conditions is documented in a PubMed study[6]. Mental Health: Significant fluctuations across demographic groups are reported by the CDC. Sleep Time: Diverse sleep behaviors are outlined in research from the Journal of Sleep Research. Physical Health: Variations due to external factors such as lifestyle choices are discussed in ScienceDirect[7]. Recognizing these outliers as plausible, we retained them in our dataset to ensure our analysis reflects the full spectrum of health profiles, enhancing the validity and comprehensiveness of our findings.

D. CRISP-DM Framework: Adapting to Heart Disease Predictive Analytics

Given the complexities and dynamic nature of heart disease prediction, particularly when considering lifestyle factors, the CRISP-DM [9] methodology stands out as particularly suitable. This framework is designed to be flexible and iterative, qualities that are essential when dealing with the evolving nature of medical data and the intricate interactions between variables. CRISP-DM's stages allow for continual refinement of models based on new insights which is crucial for translating analytical findings into practical, clinical actions.

E. Methods Overview

A. Logistic Regression Analysis Our methodology utilizes logistic regression, a robust statistical model tailored for binary classification tasks. This model hinges on probabilistic fundamentals, converting the outputs from a linear regression equation into binary outcomes (0 or 1) via a sigmoid function. Essential to the process is the selection of pertinent independent variables while eliminating those that are redundant or highly correlated, thus mitigating multicollinearity issues. To further enhance the model's stability and performance, we incorporate L2 regularization into our logistic regression framework.

For optimal parameter tuning, we deployed the Halving Grid Search CV, an efficient method for determining the best hyperparameters through a systematic exploration of specified parameter grids. This approach involved conducting 49 fits, distributed over 7 folds for each of the 7 candidate sets. The resulting optimal configuration—regularization strength (C) of 0.0016, a maximum iteration count max iter of 1000,the penalty set to ”l2”, and using 'liblinear' as the solver—significantly boosts the model's performance. This fine-tuning ensures the model not only handles the complexity well but also exhibits superior generalization capabilities, thus improving accuracy and ensuring robust operational stability.

B. Gradient Boosting Classifier Augmenting our logistic regression model, we've applied the Gradient Boosting Classifier, an advanced ensemble method that leverages the mechanics of decision trees in conjunction with gradient boosting techniques to deliver precise classification outcomes. This method effectively combines the simplicity of decision trees with the refinement of boosting, systematically enhancing

[6] https://pubmed.ncbi.nlm.nih.gov/19533240/

[7] https://www.sciencedirect.com/science/article/pii/0001457596000127

the model's performance. Each decision tree in the ensemble focuses on the errors made by its predecessors, refining its focus on the most challenging cases in an iterative fashion. The process continues, building tree after tree, until the ensemble of weak learners coalesces into a robust, unified model with considerable predictive strength, well-suited for our binary classification task even in the presence of a large dataset.

In our exploration of suitable models for our dataset, we experimented with various algorithms, including the Extra Tree Classifier, inspired by methodologies mentioned in the literature review. However, it became evident that this model prioritized accuracy over other critical metrics, leading to a suboptimal balance between recall and precision—key performance indicators for our study. Conversely, the Gradient Boosting Classifier demonstrated superior performance, particularly in optimizing recall with an impressive 80% while maintaining a solid accuracy of 71%. Though logistic regression showed commendable results, it was Gradient Boosting that ultimately excelled. Its robust handling of multicollinearity and the dataset's inherent imbalances established it as the most effective approach for our predictive analysis of heart disease prevalence based on lifestyle factors.

V. Evaluation

Evaluating model performance is crucial to determine its accuracy and effectiveness. We employ various metrics, including accuracy, F1-score, precision, and recall, to assess our models comprehensively. While accuracy measures the overall correctness of predictions, the F1-score provides a balanced view of precision (the rate of true positives among all positive predictions) and recall (the rate of true positives among actual positives). Given the initial imbalance in our dataset, relying solely on accuracy could skew results favorably towards the majority class, as noted by [11]. To mitigate this, we applied SMOTE to equalize the distribution between classes, ensuring a fair evaluation framework. In our analysis with Gradient Boosting and Logistic Regression, we focused on determining the key determinants influencing heart disease. In medical datasets, however, accuracy alone may not be indicative of a model's utility. Emphasis is placed on precision and recall, particularly reducing false negatives, to ensure critical conditions are not overlooked.

Table III shows the accuracy, f1-score, precision, and recall on Gradient Boosting Classifier.

Comparative analysis reveals that Gradient Boosting outperforms Logistic Regression, achieving an 80% recall and a 74.6% accuracy rate. This indicates a more effective model at correctly identifying true positive cases of heart disease, which is a critical measure in healthcare predictions. The evaluation emphasizes that Gradient Boosting has a higher sensitivity, ensuring a substantial proportion of actual positives are not missed (i.e., lower false negatives). As illustrated in Figure 1, the AUC-ROC curve of the Gradient Boosting model, particularly when applied to the SMOTE-enhanced dataset, demonstrates this model's superior discriminative ability to differentiate between the presence and absence of heart disease.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.757	0.56	0.78	0.51
Gradient Boosting	0.746	0.57	0.801	0.53

Table III
Evaluation Metrics

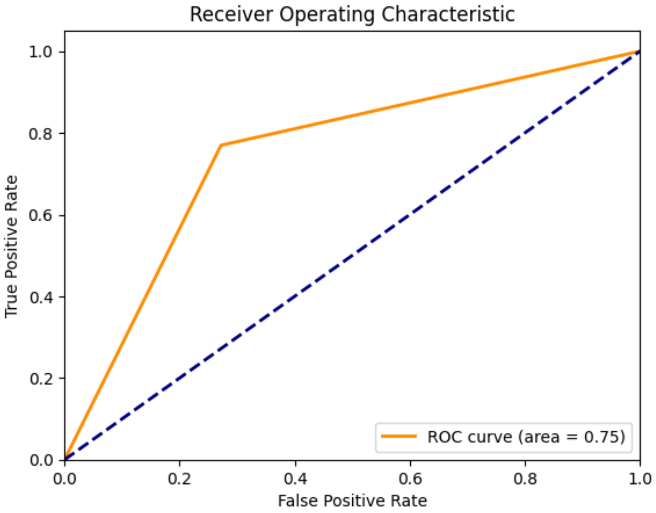

Figure 1. AUC ROC Curve from Gradient Boosting

VI. Conclusion

In conclusion, our study underscores the efficacy of Gradient Boosting as a predictive model for heart disease, particularly in addressing the challenges of an imbalanced dataset. With a notable recall of 80% and an accuracy of 74.6%, Gradient Boosting stands out for its ability to accurately identify cases of heart disease, minimizing the risk of false negatives. The model's performance, coupled with the enhanced data balance achieved through SMOTE, is further exemplified by the AUC-ROC curve presented in Figure 1. While the model demonstrates strong recall, future work could focus on enhancing precision and F1-scores to achieve a more balanced model performance. This continued refinement will not only improve the model's overall effectiveness but also set a more comprehensive benchmark for predictive models in healthcare analytics, offering a promising avenue for early detection and intervention strategies in heart disease management.

Gitlab Link to the code and dataset

REFERENCES

[1] Emelia J. Benjamin, Paul Muntner, Alvaro Alonso, Marcio S. Bittencourt, Clifton W. Callaway, April P. Carson, Alanna M. Chamberlain, Alexander R. Chang, Susan Cheng, Sandeep R. Das, Francesca N. Delling, Luc Djousse, Mitchell S.V. Elkind, Jane F. Ferguson, Myriam Fornage, Lori Chaffin Jordan, Sadiya S. Khan, Brett M. Kissela, Kristen L. Knutson, Tak W. Kwan, Daniel T. Lackland, Tené T. Lewis, Judith H. Lichtman, Chris T. Longenecker, Matthew Shane Loop, Pamela L. Lutsey, Seth S. Martin, Kunihiro Matsushita, Andrew E. Moran, Michael E. Mussolino, Martin O'Flaherty, Ambarish Pandey, Amanda M. Perak, Wayne D. Rosamond, Gregory A. Roth, Uchechukwu K.A. Sampson, Gary M. Satou, Emily B. Schroeder, Svati H. Shah, Nicole L. Spartano, Andrew Stokes, David L. Tirschwell, Connie W. Tsao, Mintu P. Turakhia, Lisa B. VanWagner, John T. Wilkins, Sally S. Wong, Salim S. Virani, and On behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. *Circulation*, 139(10), March 2019.

[2] Nadikatla Chandrasekhar and Samineni Peddakrishna. Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes*, 11(4), 2023.

[3] Matthew Davidow and David S. Matteson. Factor analysis of mixed data for anomaly detection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4):480–493, August 2022.

[4] Lampson M. Fan, Adam Collins, Li Geng, and Jian-Mei Li. Impact of unhealthy lifestyle on cardiorespiratory fitness and heart rate recovery of medical science students. *BMC Public Health*, 20(1):1012, December 2020.

[5] Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan, and Eman M. Anwar. Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method. *Complexity*, 2021:1–10, February 2021.

[6] Hamid Gholami, Aliakbar Mohammadifar, Dieu Tien Bui, and Adrian L. Collins. Mapping wind erosion hazard with regression-based machine learning algorithms. *Scientific Reports*, 10(1):20494, November 2020.

[7] Gowtham Kumar Golla, Jordan A. Carlson, Jun Huan, Jacqueline Kerr, Tarrah Mitchell, and Kelsey Borner. Developing novel machine learning algorithms to improve sedentary assessment for youth health enhancement. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 375–379, 2016.

[8] Abid Ishaq, Saima Sadiq, Muhammad Umer, Dr. Saleem Ullah, Seyedali Mirjalili, Vaibhav Rupapara, and Michele Nappi. Improving the prediction of heart failure patients' survival using smote and effective data mining techniques. *IEEE Access*, PP:1–1, 03 2021.

[9] Inna Kolyshkina and Simeon Simoff. Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach. *Frontiers in Big Data*, 4:660206, May 2021.

[10] Ying Li, Yasuto Sato, and Naohito Yamaguchi. Lifestyle Factors as Predictors of General Cardiovascular Disease: Use for Early Self-Screening. *Asia Pacific Journal of Public Health*, 26(4):414–424, July 2014.

[11] Octavio Loyola-González, José Fco. Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, and Milton García-Borroto. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing*, 175:935–947, 2016.

[12] Cynthia Yang, Egill A. Fridgeirsson, Jan A. Kors, Jenna M. Reps, and Peter R. Rijnbeek. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*, 11(1):7, January 2024.