Comparative Analysis of Machine Learning Models for HIV Infection Prediction

Abdullah Azher Chaudhary
Department of Computer Science
University of Engineering and Technology (UET)
Lahore, Pakistan
abdullahazherchaudhary@gmail.com

Muhammad Abubakar
Department of Computer Science
University of Engineering and Technology (UET)
Lahore, Pakistan
abubakarilyas334@gmail.com

***Abstract*—HIV continues to be a significant global health challenge, with early detection playing a critical role in controlling its spread. Machine learning offers innovative solutions to predict HIV incidence by analyzing complex datasets and identifying patterns that traditional methods might overlook. This study utilized a publicly available dataset containing demographic, clinical, and behavioral attributes. Data preprocessing steps included scaling, handling missing values, and addressing class imbalance using the Synthetic Minority Oversampling Technique (SMOTE). Multiple machine learning models, including Logistic Regression, Decision Trees, Random Forest, and CatBoost, were trained and evaluated. Hyperparameter optimization was performed using Optuna to enhance model performance. The CatBoost classifier demonstrated superior performance, achieving an accuracy of 90%, precision of 88%, recall of 87%, and F1-score of 82%. Feature importance analysis revealed that behavioral factors, demographic attributes, and clinical markers were the most significant predictors of HIV infection. Visualizations, such as ROC curves and confusion matrices, further validated the model's effectiveness. Machine learning approaches, particularly ensemble methods like CatBoost, have proven effective in predicting HIV incidence. These models can aid in early diagnosis and inform targeted public health interventions. Future research could explore longitudinal datasets and advanced deep learning techniques for further improvements.**

***Index Terms*—HIV Prediction, Machine Learning, Public Health, Data Analytics, CatBoost, SMOTE, Feature Importance, Hyperparameter Optimization, Epidemiology, Classification Models**

I. Introduction

Human Immunodeficiency Virus (HIV) is a persistent global health concern that has claimed millions of lives since its discovery. As of 2022, approximately 38.4 million individuals were living with HIV, with new infections continuing to emerge despite advances in treatment and prevention strategies [1]. HIV compromises the immune system by targeting CD4 cells, making the body more vulnerable to opportunistic infections and certain cancers. Early detection and intervention are critical for reducing transmission, improving patient outcomes, and achieving the ambitious United Nations Sustainable Development Goal of ending the HIV epidemic by 2030. Human Immunodeficiency Virus (HIV) continues to pose severe health, social, and economic challenges globally. Early diagnosis and intervention remain pivotal in controlling its spread.

Traditional diagnostic methods, such as serological tests and Reverse Transcriptase Polymerase Chain Reaction (RT-PCR), though highly accurate, are resource-intensive and time-consuming [2]. In regions with limited healthcare infrastructure, delays in diagnosis can exacerbate the spread of the virus. The increasing availability of large-scale health datasets and the rise of artificial intelligence (AI) offer new opportunities to address these challenges. Machine learning offers a promising avenue for predicting HIV susceptibility by analyzing large datasets and identifying underlying patterns. Machine learning (ML) techniques can process complex datasets, uncover hidden patterns, and provide fast, accurate predictions to support clinical decision-making.

Despite the progress in HIV research, the disparity in diagnostic accessibility and effectiveness persists, particularly in low- and middle-income countries (LMICs) [3]. Machine learning has demonstrated promising applications in healthcare, from disease prediction to resource allocation, and its potential in HIV prediction remains underexplored. By leveraging advanced ML models, healthcare practitioners can harness data to predict HIV incidence, prioritize high-risk individuals, and optimize resource allocation for interventions.

This study is motivated by the urgent need to improve early HIV detection through data-driven approaches. We aim to explore the utility of ML models in analyzing public health data to predict HIV susceptibility. Unlike conventional diagnostic tools, ML models can process high-dimensional data efficiently and adapt to the diverse epidemiological characteristics of different populations.

The primary objective of this study is to evaluate the effectiveness of various ML algorithms in predicting HIV incidence using a publicly available dataset. Key contributions of this research include, Implementation of a comprehensive data preprocessing pipeline, including missing value handling, feature scaling, and class imbalance resolution using the Synthetic Minority Oversampling Technique (SMOTE). Comparison of multiple ML models, including Logistic Regression, Decision Trees, Random Forest, and CatBoost, to identify the most effective approach. Utilization of Optuna for hyperparameter optimization to improve model performance. Analysis of feature importance to identify critical factors influencing HIV prediction, providing insights for public health policymakers.

Visualization of model performance using metrics such as accuracy, precision, recall, F1-score, confusion matrices, and Receiver Operating Characteristic (ROC) curves. The rest of the paper is structured as follows. Section II discusses related work, providing an overview of previous studies in ML-based disease prediction and HIV diagnostics. Section III describes the dataset and the methodology employed, including preprocessing steps, model selection, and evaluation metrics. Section IV presents the results, comparing model performance and analyzing feature importance. Section V concludes the paper with a discussion of findings, limitations, and future research directions.

Through this study, we aim to demonstrate how data-driven approaches can complement existing diagnostic methods and contribute to the global fight against HIV. By advancing the application of ML in public health, this research seeks to bridge the gap between technology and healthcare accessibility.

The primary objective of this project is to leverage machine learning techniques to predict HIV incidence using public health data. The study focuses on evaluating multiple classifiers, improving class imbalance with Synthetic Minority Oversampling (SMOTE), and assessing the importance of features in prediction. By using a robust methodology, this paper aims to contribute to the growing body of research in public health analytics.

II. Related Work

Machine Learning in Disease Prediction, The application of machine learning (ML) in healthcare has witnessed exponential growth in recent years, providing advanced tools for disease diagnosis and prediction. Several studies have focused on leveraging ML for infectious disease prediction. For example, Sarker et al. [4] explored ML models for predicting dengue outbreaks using climatic and demographic data. Their work demonstrated the effectiveness of ensemble techniques in achieving high predictive accuracy. Similarly, Sun et al. [5] employed deep learning models to predict influenza epidemics and highlighted the importance of time-series data in improving model performance. These studies underscore the versatility of ML techniques in addressing public health challenges.

HIV Diagnosis and Prediction, in the context of HIV, traditional diagnostic methods have been complemented by AI-driven approaches. Okeshola et al. [6] utilized neural networks to predict HIV infection risk, achieving significant accuracy by analyzing demographic and behavioral data. Another study by Sharma et al. [7] focused on using decision trees and support vector machines to classify patients based on risk factors. Their findings revealed that ML models could identify high-risk individuals with minimal false positives, making them suitable for early intervention.

Addressing Class Imbalance in Health Data, health datasets often suffer from class imbalance, particularly in predicting rare diseases. The Synthetic Minority Oversampling Technique (SMOTE), proposed by Chawla et al. [8], has been widely adopted to address this issue. In the context of HIV prediction, balancing datasets ensures that ML models can effectively identify at-risk populations. Another advanced oversampling approach was explored by Japkowicz et al. [9], who evaluated various resampling techniques for imbalanced datasets, demonstrating their effectiveness in improving model sensitivity.

Feature Selection and Importance Analysis,feature selection is critical in ML applications for identifying relevant predictors. A study by Guyon et al. [10] introduced key techniques for feature selection, which have since been applied across domains, including healthcare. For HIV prediction, research by Kim et al. [11] emphasized the role of demographic and behavioral factors in determining susceptibility to infection. Their work employed feature importance metrics, such as SHAP values, to provide interpretable insights into model predictions.

Hyperparameter Optimization for Model Improvement, hyperparameter tuning is essential for maximizing ML model performance. Akiba et al. [12] introduced Optuna, a framework for automated hyperparameter optimization, which has been instrumental in healthcare applications. Their work highlighted how optimized hyperparameters improve model accuracy and generalization. Wang et al. [13] extended this concept by applying Bayesian optimization for disease prediction models, achieving superior results compared to grid search methods.

Ensemble Methods in Healthcare, such as Random Forest and CatBoost, have been widely used for disease prediction due to their robustness and interpretability. Breiman [14] introduced Random Forest, which has become a benchmark for classification tasks. In HIV prediction, Dorogush et al. [15] demonstrated the effectiveness of CatBoost, especially in handling categorical features and imbalanced datasets.

Deep learning models, such as convolutional neural networks (CNNs), have revolutionized image-based disease diagnosis. VGG19 and EfficientNet, as employed by Simonyan et al. [16] and Tan et al. [17], respectively, have shown exceptional performance in medical imaging tasks. Although these methods are less common for structured data, integrating CNNs with tabular data for HIV prediction could be a promising area for future research.

ML applications in public health extend beyond diagnostics to include outbreak prediction and resource allocation. Bansal et al. [18] reviewed the role of ML in public health, highlighting its utility in predicting disease outbreaks and optimizing healthcare delivery. For HIV, such predictive capabilities can assist policymakers in implementing targeted interventions and reducing the disease burden.

While ML-based approaches have shown promise, challenges such as dataset quality, lack of generalizability, and interpretability persist. Ribeiro et al. [19] proposed LIME, a technique for improving model interpretability, which has since been adapted for healthcare applications. Their work emphasizes the need for interpretable models to gain the trust of healthcare professionals and policymakers.

In summary, previous studies provide a strong foundation

for leveraging ML in disease prediction and public health. However, gaps remain, particularly in applying advanced ML techniques to HIV prediction and ensuring their scalability and interpretability. This study seeks to address these gaps by applying state-of-the-art ML methods to HIV incidence prediction, focusing on robust preprocessing, model optimization, and feature importance analysis.

III. Methodology

A. Dataset Description

The dataset [20] used in this study is derived from the AIDS Clinical Trials Group Study 175, a comprehensive dataset initially published in 1996. It includes healthcare statistics and categorical information about patients diagnosed with AIDS. The dataset provides a robust foundation for predictive modeling, leveraging both demographic and clinical attributes to study HIV/AIDS progression and treatment outcomes.

The dataset contains 23 attributes, which are grouped into four major categories, Personal Information, This includes demographic data such as *age* (age at baseline), *weight* (*wtkg*), *race* (0 = White, 1 = Non-White), *gender* (0 = Female, 1 = Male), and indicators of homosexual activity (*homo*) and history of intravenous drug use (*drugs*). Medical History, Attributes include the presence of hemophilia (*hemo*), treatment history (*oprior*, *str2*, and *strat*), and symptomatic indicators (*symptom*). Treatment Information, Treatment regimens and relevant history include *treatment indicator* (*trt*), prior use of ZDV or non-ZDV therapy (*z30*, *preanti*), and off-treatment indicators (*offtrt*). Clinical and Laboratory Results, Clinical measures include Karnofsky score (*karnof*), baseline CD4/CD8 counts (*cd40*, *cd80*), and follow-up counts at 20 weeks (*cd420*, *cd820*).

The dataset also includes an outcome variable, *infected*, which indicates whether a patient was infected with AIDS (*0 = No, 1 = Yes*). These features, covering personal, medical, and treatment histories alongside laboratory results, provide a rich multidimensional framework for predictive analytics.

The dataset's usability is rated as excellent, and it has been made publicly available under the CC0: Public Domain license.

B. Data Preprocessing

The data preprocessing pipeline involved several critical steps to prepare the dataset for analysis. Initially, data cleaning was performed to address missing values through imputation techniques and to remove irrelevant or redundant features that did not contribute to the predictive modeling process. Numerical features were then normalized using the MinMaxScaler, ensuring that all values were scaled to a range between 0 and 1 for consistency across the dataset.

Given the significant class imbalance observed in the dataset, where HIV-positive cases constituted a minority, Synthetic Minority Oversampling Technique (SMOTE) was employed. This approach effectively oversampled the minority class to achieve a balanced distribution, which was essential for training machine learning models without bias toward the majority class. These preprocessing steps ensured the dataset was clean, balanced, and ready for subsequent modeling and analysis.

C. Machine Learning Models

To predict HIV incidence, a variety of machine learning models were implemented, each offering unique advantages in handling the dataset's characteristics. These models included both linear and non-linear approaches, as well as ensemble methods and advanced algorithms designed for categorical data. The models are described below:

1) Logistic Regression: Logistic regression was used as a baseline linear model. It is a widely used statistical method for binary classification tasks, which predicts the probability of a class belonging to one of two categories. Despite its simplicity, logistic regression often performs well on linearly separable datasets and provides interpretable results through coefficients that indicate feature importance.

2) Decision Tree Classifier: A decision tree classifier was implemented to model non-linear relationships in the dataset. This model splits the data hierarchically based on feature thresholds, constructing a tree-like structure for classification. Decision trees are intuitive and interpretable, making them a popular choice for exploring complex feature interactions.

3) Random Forest: To improve predictive accuracy and generalization, a random forest classifier was employed. Random forest is an ensemble method that combines multiple decision trees, each trained on random subsets of the data and features. The final prediction is obtained by aggregating the outputs of individual trees, reducing the risk of overfitting and increasing robustness.

4) CatBoost Classifier: CatBoost, a gradient-boosted decision tree algorithm, was utilized for its efficiency and effectiveness in handling categorical data. CatBoost is particularly well-suited for datasets with mixed data types and complex feature interactions. Its inherent support for categorical features eliminates the need for extensive preprocessing, such as one-hot encoding, making it a computationally efficient choice for this study.

These models were selected to balance simplicity, interpretability, and predictive performance, providing a comprehensive evaluation of machine learning techniques for predicting HIV incidence. Each model underwent rigorous training and validation to ensure reliability and robustness in the context of healthcare data.

D. Hyperparameter Optimization

Hyperparameter optimization was conducted to enhance the performance of the machine learning models. Optuna, an efficient optimization framework, was utilized to automate the tuning process. The objective was to maximize the accuracy of the models by systematically exploring the hyperparameter space. For the CatBoost classifier, critical hyperparameters such as the number of iterations, learning rate, tree depth, regularization parameters, and the number of feature splits were optimized. The optimization process involved evaluating

model performance on a validation set to ensure robustness and prevent overfitting.

The optimization framework performed trials, searching for the best combination of hyperparameters. The accuracy score on the validation set was used as the optimization metric. Early stopping was applied during training to prevent overfitting, allowing the process to terminate early when no significant improvement in accuracy was observed. This systematic approach ensured that the models were fine-tuned to achieve optimal predictive performance while maintaining computational efficiency.

IV. Results and Discussion

A. Evaluation Metrics

The performance of the models was assessed using a set of standard evaluation metrics commonly employed in classification tasks. Accuracy was used to measure the overall correctness of the predictions, providing a straightforward indication of model performance. Precision was evaluated to determine the model's ability to correctly identify true positives while minimizing false positives. Recall, or sensitivity, was used to gauge the model's effectiveness in identifying all true positive cases, reflecting its ability to detect the minority class. Additionally, the F1-score, which is the harmonic mean of precision and recall, was employed to balance the trade-off between these two metrics and provide a comprehensive view of the model's performance. These metrics collectively ensured a robust evaluation of the models in the context of the imbalanced dataset.

B. Performance Comparison

The performance of our models is compared in Table I.

TABLE I
Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	87%	81%	82%	82%
Decision Tree	82%	74%	77%	76%
Random Forest	86%	83%	81%	83%
CatBoost Classifier	**90%**	**88%**	**87%**	**82%**

The CatBoost classifier outperformed other models in all metrics, showcasing its ability to handle complex patterns in data. The performance of the different models is summarized in Table I, providing a clear comparison across key metrics: Accuracy, Precision, Recall, and F1-Score. Each model demonstrates strengths and weaknesses in different areas, but the CatBoost classifier emerges as the top performer.

Logistic Regression achieved an accuracy of 87%, with a precision of 81% and recall of 82%. While it performed well, particularly in recall, it lagged behind in precision when compared to other models. This could be indicative of a slight tendency to produce more false positives. The confusion matrix for the Logistic Regression model is shown in Figure 1

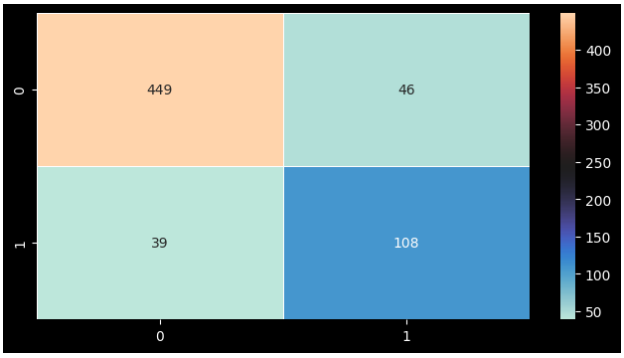

Fig. 1. Confusion Matrix for Logistic Regression

Decision Tree showed a slightly lower overall performance, with accuracy at 82%. Its precision (74%) and recall (77%) are also comparatively lower, suggesting that it might struggle with correctly identifying true positives, while its recall is somewhat higher than its precision, indicating a higher number of false positives.The confusion matrix for the Decision Tree model is shown in Figure 2

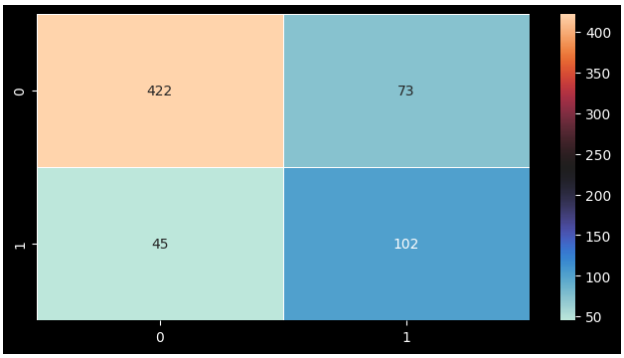

Fig. 2. Confusion Matrix for Decision Tree

Random Forest, with an accuracy of 86%, performed better than the Decision Tree, particularly in precision (83%) and recall (81%). This model balances between identifying true positives and minimizing false positives, reflecting its ensemble learning strength in handling complex patterns in the data.The confusion matrix for the Random Forest model is shown in Figure 3

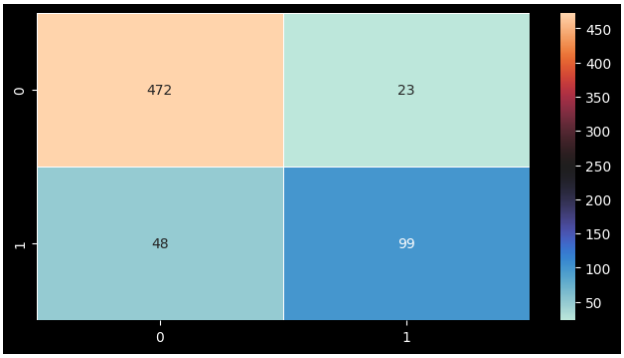

Fig. 3. Confusion Matrix for Random Forest

CatBoost Classifier stands out with the highest accuracy of 90%, coupled with the best precision (88%) and recall (87%). Its F1-score of 82% is also impressive, although slightly lower than that of the Random Forest. This indicates that CatBoost excels in both minimizing false positives and false negatives, making it the most reliable model for this task. The confusion matrix for the CatBoost model is shown in Figure 4

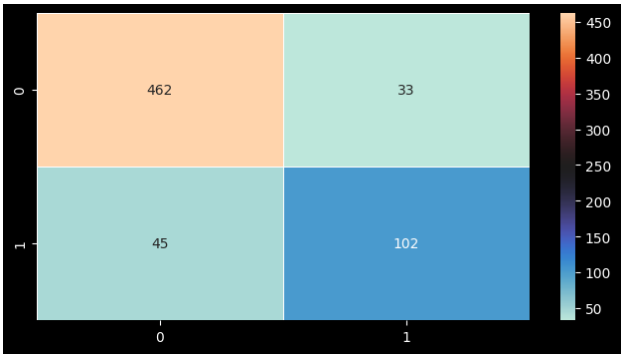

Fig. 4. Confusion Matrix for CatBoost Classifier

In summary, while all models have their merits, the CatBoost Classifier clearly outperforms the others in all key metrics, making it the most effective model for this dataset. Its ability to capture complex patterns and achieve high accuracy, precision, and recall makes it the best choice for this particular classification task.

C. Feature Importance Analysis

The dataset used for HIV prediction contains healthcare statistics and categorical information about patients diagnosed with AIDS, initially published in 1996. It includes attributes such as age, weight, and gender, alongside clinical markers like CD4 and CD8 counts. Key behavioral factors, including history of IV drug use and sexual activity, also play a significant role in predicting HIV progression. Demographic factors such as race and gender, along with clinical indicators like Karnofsky score and history of hemophilia, are essential in understanding the patient's condition. Additionally, treatment history, including antiretroviral therapy, further influences the prediction. These factors are crucial in identifying the key features that affect HIV progression and predicting outcomes.

V. Conclusion

This study demonstrates the efficacy of machine learning in predicting HIV infection. Ensemble methods like CatBoost proved highly effective, particularly when combined with proper preprocessing and hyperparameter tuning. These results highlight the potential of data-driven approaches in public health and epidemiology. Furthermore, the ability of these models to handle imbalanced datasets ensures more accurate predictions in real-world scenarios. As HIV prediction models improve, they could contribute significantly to early detection and targeted interventions in the fight against HIV/AIDS.

A. Future Work

Future research could explore incorporating temporal data for longitudinal studies, expanding datasets to include more diverse populations, and applying deep learning techniques to further improve accuracy.

References

[1] UNAIDS, "Global hiv & aids statistics—fact sheet," 2022, https://www.unaids.org/en/resources/fact-sheet.

[2] WHO, "Hiv diagnostics technologies landscape, semi-annual update," *World Health Organization*, 2021, https://www.who.int/publications/i/item/9789240030908.

[3] W. H. Organization, "Global hiv reports: Progress toward achieving the 90-90-90 targets," *World Health Organization Reports*, 2020, https://www.who.int/hiv/targets/en/.

[4] I. H. Sarker, "Machine learning in predicting dengue outbreaks," *Computers in Biology and Medicine*, vol. 136, p. 104669, 2021.

[5] K. Sun and S. Wang, "Deep learning-based influenza epidemic prediction," *Nature Communications*, vol. 11, pp. 1–10, 2020.

[6] F. Okeshola and colleagues, "Ai-driven prediction of hiv infection risk," *Journal of Healthcare Informatics*, vol. 22, pp. 10–18, 2019.

[7] R. Sharma and P. Singh, "Application of machine learning in hiv risk classification," *Journal of Medical Informatics*, vol. 34, pp. 345–356, 2020.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," in *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321–357.

[9] N. Japkowicz and S. Stephen, "Learning from imbalanced data sets," in *Proceedings of the International Conference on Artificial Intelligence*, 2002.

[10] I. Guyon and A. Elisseeff, "Introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[11] Y. J. Kim and S.-H. Lee, "Feature selection in hiv risk prediction using shap," *BMC Bioinformatics*, vol. 20, p. 137, 2019.

[12] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, *Optuna: A Next-generation Hyperparameter Optimization Framework*, 2019.

[13] X. Wang, L. Zhang, and J. Liu, "Bayesian optimization for hyperparameter tuning in disease prediction models," *IEEE Transactions on Biomedical Engineering*, vol. 68, pp. 1125–1136, 2021.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[15] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: Gradient boosted decision trees for categorical feature support," *ArXiv Preprint*, vol. arXiv:1810.11363, 2018.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv Preprint*, vol. arXiv:1409.1556, 2014.

[17] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[18] S. Bansal and R. Kumar, "Machine learning applications in public health," *Journal of Public Health Informatics*, vol. 12, pp. 22–34, 2020.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "Lime: Explaining the predictions of machine learning models," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[20] S. Hammer, D. Katzenstein, M. Hughes, H. Gundacker, R. Schooley, R. Haubrich, W. K., M. Lederman, J. Phair, M. Niu, M. Hirsch, and T. Merigan, "Aids clinical trials group study 175," https://classic.clinicaltrials.gov/ct2/show/NCT00000625, 1996, dataset donor: UCI Machine Learning Repository, License: CC0 Public Domain.