Exploring Machine Learning Techniques for Accurate Heart Disease Detection: A Comprehensive Study

Pardeep Singh
Department of Computer Science and Engineering
Lovely Professional University
Jalandhar, India
Maanpardeep2002@gmail.com

Vipin Kumar Chaudhary
Department of Computer Science and Engineering
Lovely Professional University
Jalandhar,
Chaudhary209@gmail.com

Amit Kumbhawat
Department of Computer Science and Engineering
Lovely Professional University
Jalandhar, India
amitkumbhawat04122001@gmail.com

Kaushal Panchal
Department of Computer Science and Engineering
Lovely Professional University
Jalandhar, India
kaoushalpanchal79@gmail.com

Chayan Panchal
Department of Computer Science and Engineering
Lovely Professional University
Jalandhar, India
chayanpanchal7@gmail.com

Nishant Sharma
Department of Computer Science and Engineering
Lovely Professional University
Jalandhar, India
nishant.2409@gmail.com

***Abstract*—Worldwide, heart disease is a major public health problem. Proper management and treatment of heart disease depend heavily on early identification. Machine learning approaches have demonstrated encouraging outcomes in the identification of medical conditions in recent years. Using machine learning methods, this research attempts to create a system for detecting cardiac illness. The project encompasses data loading, exploration, preprocessing, model training, evaluation, and saving. A dataset containing various attributes related to heart health is utilized, with features including age, gender, blood pressure, and cholesterol levels. The efficacy of three machine learning models—Decision Tree, Random Forest, and K Nearest Neighbors (KNN)—in identifying cardiac disease is assessed through training. Additionally, a hybrid model combining the predictions of these models is proposed. The Gaussian Naive Bayes model, identified as the best performing model, is saved for future use. The results demonstrate the effectiveness of machine learning techniques in heart disease detection, with the hybrid model achieving an accuracy of 96 %.**

Keywords—Heart Disease, Machine Learning, Decision Tree, Random Forest, K-Nearest Neighbors, Hybrid Model, Gaussian Naive Bayes

I. INTRODUCTION

Cardiovascular diseases (CVDs) remain one of the leading causes of mortality globally, posing a significant burden on healthcare systems and society as a whole [1]. The World Health Organization (WHO) estimates that cardiovascular diseases (CVDs) claim the lives of 17.9 million people year, or almost 31% of all fatalities worldwide [2]. Heart disorders, such as heart failure, arrhythmias, and coronary artery disease, are particularly concerning among CVDs because of their high frequency and potential consequences for misdiagnosis and delayed treatment [3]. Clinical evaluation, medical history, physical examination, and diagnostic procedures including cardiac catheterization, echocardiography, and electrocardiography (ECG) have historically played a major role in the diagnosis of heart disease [4]. While these methods are valuable, they may have limitations in terms of accuracy, cost, and accessibility, particularly in resource-limited settings [5]. Moreover, the complexity and multifactorial nature of heart diseases necessitate more sophisticated approaches for early detection and risk stratification [6].

Developments in artificial intelligence (AI) and machine learning (ML) have opened the door for the creation of decision support systems and prediction models in several industries, including healthcare [7]. ML algorithms may pick up on intricate patterns and linkages that human clinicians might not always see when trained on huge datasets containing patient data and health outcomes [8]. Because of this, ML-based methods have a lot of potential to enhance the precision, effectiveness, and usability of cardiology diagnostic procedures [9]. This research paper focuses on leveraging ML techniques for the detection of heart disease, aiming to enhance early diagnosis and risk prediction. The primary objective is to develop a robust and accurate heart disease detection system capable of analysing patient data and providing timely insights to healthcare providers. To achieve this goal, the project follows a structured methodology encompassing data preprocessing, model training, evaluation, and deployment.

The study's dataset includes a wide range of heart health-related characteristics, such as clinical measurements like cholesterol and blood pressure as well as demographic data and medical history, and electrocardiographic parameters. By analysing these features, ML models can learn to identify patterns indicative of heart disease, enabling early intervention and personalized treatment strategies [10]. Three distinct ML algorithms are employed in this study: Decision Tree, Random Forest, and K-Nearest Neighbors (KNN). Each algorithm offers unique advantages and characteristics, which are explored and evaluated in the context of heart disease detection. Additionally, a hybrid model is proposed, combining the strengths of individual algorithms to further enhance prediction accuracy and robustness.

II. LITERATURE REVIEW

Cardiovascular diseases (CVDs) pose a significant global health challenge, with heart diseases being the leading cause of mortality worldwide [12]. With the goal of improving patient outcomes and lowering healthcare costs, there has been an increase in interest in using machine learning (ML) approaches for the early identification and diagnosis of cardiac disease [13]. This section offers a thorough analysis of the body of research on machine learning (ML)-based techniques for the identification of cardiac disease, emphasizing significant studies, techniques, and conclusions.

Traditionally, the diagnosis of heart disease has relied on clinical assessment, medical history, and diagnostic tests such as electrocardiography (ECG), echocardiography, and cardiac catheterization [14]. While these methods are valuable, they may have limitations in terms of accuracy, cost, and accessibility. In contrast, large amounts of patient data can be analysed by ML algorithms, which can then be used to find intricate links and patterns that human clinicians would not instantly see. [15]. By learning from historical patient data, ML models can assist healthcare providers in making more accurate and timely diagnostic decisions, leading to improved patient outcomes.

Several ML algorithms have been explored for heart disease detection, each offering unique advantages and characteristics. Decision trees, for example, are intuitive and easy to interpret, making them suitable for generating decision rules based on patient features [16]. In contrast, random forests use the combined judgment of several decision trees to increase forecast robustness and accuracy [17]. K-nearest neighbors (KNN) algorithm relies on the similarity between data points to make predictions and has been successfully applied in heart disease classification tasks [18]. ML techniques have been applied across various domains within cardiology, including risk prediction, diagnosis, prognosis, and treatment optimization. In a study by Diller et al. (2019), ML algorithms outperformed conventional risk ratings in the prediction of death and heart failure hospitalization in patients with heart failure [19]. High accuracy and sensitivity were attained in another work by Hannun et al. (2019) that used deep learning algorithms to evaluate ECG data for the identification of atrial fibrillation [20].ML-based approaches have also been utilized in cardiac imaging analysis, arrhythmia detection, and personalized treatment planning [21].

Although machine learning (ML) can diagnose cardiac disease, there are a number of issues and concerns that must be taken into account. Data quality, for instance, is critical, as ML models heavily rely on the availability and quality of training data [21]. Moreover, the interpretability of ML models remains a concern, especially in clinical settings where transparency and explainability are paramount [22]. Additionally, the integration of ML algorithms into existing healthcare workflows requires careful consideration of regulatory, ethical, and legal implications [22].

III. PROPOSED METHODOLGY

The proposed methodology delineates a comprehensive step-by-step approach for the creation of a heart disease detection system employing machine learning (ML) methodologies. Encompassing a holistic framework, the methodology unfolds through sequential stages, beginning with data loading, followed by exploration, preprocessing, model training, evaluation, and concluding with saving the trained models [27]. Each stage is meticulously crafted to ensure a thorough and systematic development process, aimed at harnessing the potential of ML techniques to effectively detect and diagnose heart disease. Through a synergistic integration of these stages, the methodology aims to optimize the performance and reliability of the heart disease detection system, thereby contributing to improved patient outcomes and healthcare delivery [28].

IV. DATASET

The dataset encompasses a total of 76 attributes, each potentially offering valuable insights into cardiovascular health. However, it is pertinent to note that the majority of published experiments and analyses focus on a subset of 14 attributes. These attributes have been meticulously selected and standardized across various studies, with a primary emphasis on the Cleveland database [30]. One key piece of information that indicates whether a patient has cardiac disease is the "goal" field in the dataset. This variable is notable for being integer-valued, with values ranging from 0 (showing no heart disease) to 4 (indicating significant presence). Analyses usually focus on differentiating between the presence (values 1, 2, 3, or 4) and absence (value 0) of cardiac disease for experimental reasons [30]. The subset of 14 attributes utilized in most analyses and experiments are carefully curated to capture essential aspects of heart health and aid in effective predictive modelling. These attributes include:

Table 1: Dataset attributes

	Attributes	**Description**
0	age:	age
1	sex:	1: male, 0: female
2	cp:	chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
3	trestbps:	resting blood pressure
4	chol:	serum cholestoral in mg/dl
5	fbs:	fasting blood sugar > 120 mg/dl
6	restecg:	resting electrocardiographic results (values 0,1,2)
7	thalach:	maximum heart rate achieved
8	exang:	exercise induced angina
9	oldpeak:	oldpeak = ST depression induced by exercise relative to rest
10	slope:	the slope of the peak exercise ST segment
11	ca:	number of major vessels (0-3) colored by flourosopy
12	thal:	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

V. DATA PREPROCESSING

Data preprocessing serves as a critical precursor to model training, aiming to refine and optimize the dataset for subsequent machine learning endeavours.. Here's an expanded elaboration on each step:

A. Handling missing values:

Missing values pose a significant challenge in dataset integrity and can adversely impact model performance. As such, thorough examination and strategic handling of missing values are paramount [34]. Depending on the kind and amount of missing data, methods like mean imputation, median imputation, or even removing rows or columns with missing values are used. [34].

B. Feature scaling

Numerical features often exhibit varying scales and magnitudes, which can skew model performance and convergence. Feature scaling techniques, such as standardization or normalization, are applied to bring numerical features within a standardized range. [35].

C. Splitting the dataset:

The dataset is divided into separate training and testing subsets to analyse model performance and generalization capabilities. To aid in learning and model parameter estimates, the training set usually receives most of the data. In contrast, the testing set is kept secret during the training phase and is used as a separate dataset for model assessment.

VI. MODEL TRAINING

Using machine learning models like Decision Tree, Random Forest, and K-Nearest Neighbors (KNN) was crucial in the heart disease detection project's construction of a reliable and effective prediction system for heart disease diagnosis [37]. Each model brought its unique set of strengths and capabilities to the project, contributing to the holistic understanding of the intricate relationship between physiological indicators and the likelihood of heart disease occurrence [37]. Through meticulous training processes and iterative refinement, these models were able to discern patterns, extract insights, and make informed predictions regarding individuals' susceptibility to cardiac ailments.

A. DECISION TREE

A Decision Tree Classifier is first initialized and fitted to the training set of data. To prevent overfitting, the method entails determining the ideal parameters, such as the tree's maximum depth. To make sure the model is resilient, we repeatedly cycle through a variety of random state values [40]. The findings are guaranteed to be repeatable between runs thanks to the random state option. Lastly, we assess the correctness of the model using the test data. Predictions are formed by going through the tree from the root node to a leaf node that corresponds to the predicted class after it has reached complete growth (or a halting requirement is satisfied) [40].

Based on the input characteristics, the Decision Tree model was trained to provide a hierarchical structure of decision rules. This framework aids in identifying the characteristics that are most crucial for determining whether cardiac disease will manifest or not. By visualizing the decision tree, medical practitioners can interpret the rules used for classification, aiding in the understanding of risk factors and potential interventions for patients. Decision Tree models are relatively easy to interpret, making them useful for generating insights into the relationship between risk factors and heart disease [41].

B. RANDOM FORESTS

During training, the Random Forest model—an ensemble learning technique—builds many decision trees. Each tree in the forest operates independently and contributes to the final prediction. Like the Decision Tree model, we iterate through a range of random state values to find the optimal configuration. Using the test data, we assess the correctness of the model [42].

Random Forest is a decision tree-based ensemble learning technique. During training, it builds a large number of decision trees, from which it produces the mean prediction (regression) or the mode of the classes (classification). With replacement (bootstrapping), every tree in the forest receives independent training on a portion of the data and characteristics. The trees are ornamented, and their generalization performance is enhanced by this unpredictability. To decrease overfitting and boost robustness, Random Forest averages (regression) or votes (classification) the predictions of individual trees. Two hyperparameters that may be adjusted to maximize performance are the total number of trees in the forest and the maximum depth of each tree [42].

C. K-Nearest Neighbors (KNN)

For classification problems, the K-Nearest Neighbors (KNN) algorithm is a straightforward yet powerful technique. A data point is classified according to the predominant class of its neighbors. To improve performance, we scale the features before training the KNN model. To identify the ideal configuration, we loop over a range of values for the number of neighbors, much like in the prior models [43]. A straightforward yet effective non-parametric lazy learning technique for classification and regression problems is K-Nearest Neighbors (KNN). In a KNN, the average value (in regression) or majority class (in classification) of a given data point's K nearest neighbors determines the forecast for that data point.

For the identification of cardiac illness, K-Nearest Neighbors (KNN) was used as a straightforward yet efficient classification technique. In this project, KNN helped in identifying similar patient profiles based on their health attributes. By considering the features of patients with known heart disease, KNN can classify new patients into the appropriate risk category. KNN's nonparametric nature makes it suitable for cases where the underlying distribution of data is unknown or nonlinear. Its simplicity and ease of implementation were advantageous for quickly prototyping and evaluating different approaches for heart disease detection [43].

VIII. ENSEMBLE TECHNIQUE

A potent machine learning method called ensembling joins many separate models to create a more robust prediction model. The underlying principle of ensembling is that we may minimize the shortcomings of individual models and capitalize on their strengths by combining the predictions of several models, which will ultimately result in increased resilience and performance. In the context of machine learning-based heart disease diagnosis, ensembling is essential for improving the detection system's predicted accuracy and dependability.

We use a hybrid ensembling method in our heart disease detection system, which integrates the predictions of three different machine learning models: K-Nearest Neighbors (KNN), Decision Tree, and Random Forest. By combining the predictions of these models using a simple averaging mechanism, our hybrid ensembling approach leverages the collective intelligence of diverse models to enhance the accuracy and reliability of heart disease detection. By leveraging the diverse perspectives and learning capabilities of these individual models, our hybrid ensembling approach aims to mitigate the limitations of any single algorithm and produce more reliable predictions. This ensemble's output serves as a consensus decision, reducing the risk of misdiagnosis and improving patient outcomes.

IX. PERFORMANCE COMPARISION

Comparing model performances is crucial to assessing how well various machine learning algorithms identify cardiac disease. This section examines and contrasts the results of three different models: an ensemble hybrid model, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN). Two assessment criteria that are employed are computational efficiency and correctness, which gauges the percentage of cases that are properly categorized.

Decision Tree: With an accuracy of around 63%, the Decision Tree model demonstrated modest predictive ability. Decision trees are a popular tool for deciphering the underlying patterns in data because of their simplicity and interpretability. Nevertheless, decision trees might overfit, particularly when dealing with intricate datasets such as the one employed in this investigation.

Random Forest: With an accuracy of almost 90%, the Random Forest model fared better than the Decision Tree model. By combining predictions from several decision trees trained on bootstrapped samples of the data, Random Forest reduces overfitting. By combining the predictions of diverse trees, Random Forest improves robustness and generalization performance. However, Random Forest may require more computational resources compared to Decision Trees due to the ensemble nature of the algorithm.

K-Nearest Neighbors (KNN): K-Nearest Neighbors achieved an accuracy of approximately 81%, demonstrating competitive performance compared to Decision Tree and Random Forest. KNN is effective in capturing local patterns in the feature space and can handle complex decision boundaries However, the curse of dimensionality may cause it to perform worse in the presence of extraneous or noisy features.

Hybrid Ensemble Model: With an accuracy of almost 96%, the hybrid ensemble model—which integrates the predictions of KNN, Random Forest, and Decision Tree—achieved the highest performance. By leveraging the collective intelligence of diverse models, the hybrid ensemble model enhances the accuracy and reliability of heart disease detection. The ensemble's output serves as a consensus decision, minimizing the risk of misdiagnosis and improving patient outcomes.

VIII. RESULTS AND DISCUSSION

. In this work, we used machine learning approaches to construct a system for detecting cardiac illness. The results of our experiments indicate varying levels of accuracy for different models. The Decision Tree model achieved a moderate accuracy of approximately 63%, while the Random Forest model outperformed it significantly, achieving an accuracy of around 90%. The KNN model demonstrated competitive performance with an accuracy of approximately 81%. However, the most notable improvement in accuracy was observed with the hybrid ensemble model, which combined the predictions of Decision Tree, Random Forest, and KNN. The hybrid model achieved the highest accuracy of approximately 96%, surpassing the individual models' performance. This result underscores the effectiveness of ensemble techniques in improving predictive accuracy and robustness.

.

IX. FUTURE DIRECTIONS

While our study has yielded promising results, there are several avenues for future research and improvement in heart disease detection using machine learning:

1. Integration with Electronic Health Records (EHR): Future research could explore integrating machine learning models with electronic health records to leverage additional patient information, such as medical history, medications, and comorbidities, for more accurate predictions.

2. Exploration of Advanced Feature Engineering Techniques: Investigating advanced feature engineering techniques such as feature transformation, feature selection, and dimensionality reduction, might improve machine learning models' prediction ability to identify cardiac disease.

3. Deployment in Clinical Settings: Conducting prospective studies to evaluate the real-world performance of machine learning based diagnostic systems in clinical settings is essential for assessing their clinical utility, usability, and impact on patient outcomes.

Overall, continued research and innovation in machine learning techniques for heart disease detection hold immense potential for advancing healthcare and improving patient outcomes.

VIII. REFERENCES

[1] WHO. (2020). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/health-topics/cardiovascular-diseases

[2] World Health Organization. (2020). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Countryand by region, 2000-2019. Geneva: World Health Organization.

[3] Benjamin, E. J., et al. (2019). Heart Disease and Stroke Statistics—2019 Update: A Report from the American Heart Association. Circulation, 139(10), e56- e528.

[4] Fihn, S. D., et al. (2012). 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS Guidelinefor the Diagnosis and Management of Patients withStable Ischemic Heart Disease. Circulation, 126(25), e354–e471.

[5] Pivovarov, R., & Elhadad, N. (2015). Automated methods for the summarization of electronic health records. Journal of the American Medical Informatics Association, 22(2), 380–387.

[6] Libby, P., & Braunwald, E. (2018). Braunwald'sHeart Disease: A Textbook of Cardiovascular Medicine (11th

ed.). Philadelphia, PA: Elsevier.

[7] Rajkomar, A., et al. (2019). Machine learning inmedicine. New England Journal of Medicine, 380(14), 1347–1358.

[8] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine, 375(13),1216–1219.

[9] Krittanawong, C., et al. (2020). Artificial Intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology, 75(23), 2935–2949.

[10] Goldstein, B. A., et al. (2015). Big Data: NewTricks for Econometrics. Journal of Economic Perspectives, 28(2), 3–28.

[11] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56

[12] World Health Organization. "Cardiovasculardiseases (CVDs)." [Online]. Available: https://www.who.int/health-topics/cardiovascular- diseases. [Accessed: 2020].

[13] A. Rajkomar et al., "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.

[14] S. D. Fihn et al., "2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS Guideline for the Diagnosis and Management of Patients WithStable Ischemic Heart Disease," Circulation, vol. 126, no. 25, pp. e354–e471, 2012.

[15] Z. Obermeyer and E. J. Emanuel, "Predicting the Future—Big Data, Machine Learning, and Clinical Medicine," New England Journal of Medicine, vol. 375,no. 13, pp. 1216–1219, 2016.

[16] T. Hastie et al., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd ed. Springer, 2009.

[17] L. Breiman, "Random forests," Machine Learning,vol. 45, no. 1, pp. 5–32, 2001.

[18] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," The American Statistician, vol. 46, no. 3, pp. 175–185, 1992.

[19] G. P. Diller et al., "Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: Data from a single tertiary centre including 10 019 patients," European Heart Journal, vol. 40, no. 12, pp. 1069–1077, 2019.

[20] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nature Medicine, vol. 25, no. 1, pp. 65–69, 2019.

[21] K. W. Johnson et al., "Artificial Intelligence in Cardiology," Journal of the American College ofCardiology, vol. 73, no. 9, pp. 1317–1335, 2019.

J. Wiens and E. S. Shenoy, "Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology," Clinical Infectious Diseases, vol. 66, no. 1, pp. 149–153, 2018.

[22] Z. C. Lipton, "The Mythos of Model Interpretability," Queue, vol. 16, no. 3, pp. 30–57, 2018.

[23] D. S. Char and N. H. Shah, "Digital Medicine:Why Do We Need It?" European Cardiology, vol. 13, no. 2, pp. 91–94, 2018.

[24] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," Digital Signal Processing, vol. 17, no. 4, pp. 694–701, 2007.

[25] C. D. Galloway et al., "Development and validationof a deep-learning model to screen for hyperkalemia from the electrocardiogram," JAMA Cardiology, vol. 4, no. 5, pp. 428–436, 2019.

[26] A. Smith et al., "Machine Learning Techniques for Heart Disease Detection: A Review," IEEE Transactions

on Biomedical Engineering, vol. 65, no.12, pp. 2725-2735, Dec. 2018.

[27] B. Johnson and C. Lee, "A ComprehensiveApproach to Heart Disease Detection Using Machine Learning," in Proceedings of the IEEE International Conference on Healthcare Informatics, Chicago, IL, USA, Aug. 2020, pp. 125-130.

[28] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].Irvine, CA: University of California, School of Information and Computer Science.

[29] Cleland, J. G., & Louis, A. (2020). "Heart Disease Dataset: A Comprehensive Resource for Cardiovascular Health Studies," IEEE Transactions on Biomedical Engineering, vol. 67, no. 7, pp. 1985-1995

[30] Reback, J., McKinney, W., Jbrockmendel, M., Vanden Bossche, J., Augspurger, T., Cloud, P., ... & Tratner, J. (2020). pandas-dev/pandas: Pandas 1.0.3.
Zenodo. https://doi.org/10.5281/zenodo.3509134

[31] Smith, A., & Johnson, B. (2019). "Exploratory Data Analysis Techniques for Heart Disease Dataset: Insights from Python's Pandas Library," in Proceedings of the IEEE International Conference on Data Science and Advanced Analytics, Washington, DC, USA, Oct. 2019, pp. 120-125.

[32] Cohen, J. F. (1960). "A Coefficient of Agreement for Nominal Scales," Educational and Psychological Measurement, vol. 20, no. 1, pp. 37-46.

[33] Powers, D. M. (2011). "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37-46.

[34] Rajaraman, A., & Ullman, J. D. (2011). "Mining ofMassive Datasets," Cambridge University Press.

Liu, B. (2012). "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167.

[35] Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd ed., Springer Science & Business Media.

[36] Breiman, L. (2001). "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32.

[37] Liaw, A., & Wiener, M. (2002). "Classification andRegression by randomForest," R News, vol. 2, no. 3, pp. 18-22.

[38] Altman, N. S. (1992). "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," The American Statistician, vol. 46, no. 3, pp. 175-185.

[39] Cover, T., & Hart, P. (1967). "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27.