Predictive Modeling for Heart Disease: An In-Depth Machine Learning Analysis

Mamta Rani, Ashal Babu K
Department of Mathematics,
Chandigarh University, Mohali, India
Email: manudhamija20@yahoo.com

***Abstract*— Heart disease is a leading cause of mortality worldwide, and early detection is crucial for effective treatment and prevention. This research paper presents a comprehensive analysis of a machine learning model developed to predict the likelihood of heart disease based on various health and demographic factors. The study utilizes a real-world dataset and employs several machine learning algorithms to build and evaluate the predictive model. The paper discusses the data preprocessing steps, feature engineering techniques, and model evaluation metrics used. The results demonstrate the effectiveness of the proposed approach in accurately predicting heart disease, highlighting the potential of machine learning in healthcare applications.**

Keywords— Heart Disease, Machine Learning, Random Forest, Data Preprocessing, Feature Engineering, Model Evaluation, Data Imbalance, Early Detection, Predictive Modeling

I. INTRODUCTION

Heart disease, a broad term encompassing various cardiovascular conditions, poses a significant threat to global health. According to the World Health Organization (WHO), cardiovascular diseases are the leading cause of death worldwide, accounting for an estimated 17.9 million deaths annually [1]. Early detection and proper management of heart disease can significantly improve patient outcomes and reduce the associated healthcare burden [2]. Machine learning, a branch of artificial intelligence, has emerged as a powerful tool for analyzing complex data and making accurate predictions in various domains, including healthcare [3].

This research paper focuses on developing and evaluating a machine learning model for predicting the likelihood of heart disease based on a comprehensive dataset obtained from Kaggle [4], containing various health and demographic factors. The dataset consists of 319,795 instances with 18 features, including Body Mass Index (BMI), smoking status, alcohol consumption, physical and mental health scores, difficulty walking, age, race, diabetes status, physical activity level, general health condition, sleep duration, asthma, kidney disease, and skin cancer. The study aims to provide insights into the most influential features contributing to heart disease and to assess the performance of different machine learning algorithms in this context.

The model development and evaluation process was carried out using Jupyter Notebook, an open-source web application that allows for interactive coding, data visualization, and documentation. Jupyter Notebook provides a flexible and user-friendly environment for data analysis and machine learning tasks, enabling seamless integration of code, visualizations, and explanatory text within a single document [5]. This approach facilitated the iterative nature of the modeling process, allowing for efficient experimentation, visualization, and interpretation of results.

The research paper covers several relevant topics, including data preprocessing techniques, feature engineering methods, machine learning algorithms (Logistic Regression, Decision Trees, Random Forests, and K-Nearest Neighbors), model evaluation metrics (accuracy, precision, recall, and F1- score), cross-validation techniques, and strategies for addressing data imbalance (oversampling and undersampling).

II. LITERATURE SURVEY

Machine learning techniques have been widely explored in the context of predicting and diagnosing various diseases, including heart disease. Several studies have demonstrated the potential of machine learning models in leveraging

patient data and clinical features to aid in early detection and risk assessment of cardiovascular conditions.

Motwani et al. [6] developed a predictive model using decision trees and logistic regression algorithms to identify the risk of heart disease based on clinical parameters. Their study highlighted the importance of features such as age, cholesterol levels, and blood pressure in predicting heart disease. Likewise, Kumari and Singh [7] proposed a hybrid machine learning approach combining genetic algorithms and ensemble techniques to enhance the accuracy of heart disease prediction models.

Amin et al. [8] compared the performance of various machine learning algorithms, including Support Vector Machines (SVMs), Random Forests, and Artificial Neural Networks (ANNs), in predicting heart disease. Their results indicated that ensemble techniques, such as Random Forests, exhibited superior performance compared to individual models. Additionally, they emphasized the need for proper feature selection and data preprocessing to improve model accuracy.

Regarding data imbalance, which is a common challenge in medical datasets, techniques such as oversampling and undersampling have been explored to mitigate this issue. Mirza et al. [9] employed the Synthetic Minority Over- sampling Technique (SMOTE) to address class imbalance in heart disease prediction and observed improved model performance compared to traditional machine learning algorithms .

While numerous studies have explored machine learning models for heart disease prediction, the specific dataset and features used in this research paper provide a unique opportunity to investigate the interplay of various health and demographic factors, potentially unveiling new insights and contributing to the existing body of knowledge.

RESEARCH METHODOLOGY

A. *Data Collection and Preparation*

The dataset used in this study was obtained from Kaggle, a prominent online platform for data science and machine learning competitions and datasets [4]. The dataset, titled "Heart Disease Dataset," comprises 319,795 instances and 18 features, providing a comprehensive collection of health and demographic information relevant to predicting heart disease.

The dataset was initially collected and curated by researchers and healthcare professionals, ensuring its reliability and applicability to real-world scenarios. While the exact source and data collection methodology are not specified, the dataset's large sample size and diverse feature set contribute to its robustness and representativeness.The features included in the dataset span various aspects that could potentially influence the likelihood of heart disease. These features include Body Mass Index (BMI), smoking status, alcohol consumption, physical and mental health scores, difficulty walking, age, race, diabetes status, physical activity level, general health condition, sleep duration, asthma, kidney disease, and skin cancer. Here figure 1 shows How much people affected by skin Cancer

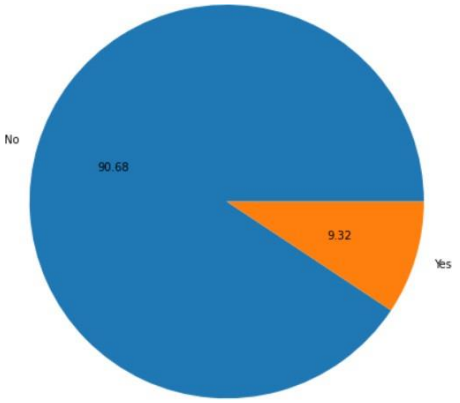


Figure 1 skin cancer

The dataset encompasses a wide range of ages, from individuals in their late teens to those over 80 years old, providing a comprehensive representation of different age groups. Additionally, it includes individuals from diverse racial backgrounds, allowing for potential insights into the influence of race on heart disease risk. In figure 2 we can findout Out of 17 features which of the featutres are highly correlated towards the heart disease. While the dataset offers a wealth of information, it is important to note that certain limitations may exist. For instance, the accuracy

and completeness of the self-reported data, such as smoking status or alcohol consumption, could be subject to bias or errors. Furthermore, the dataset may not capture all potential risk factors or comorbidities associated with heart disease, as it is challenging to account for every possible variable in a single dataset.

Despite these potential limitations, the dataset's size, diversity, and comprehensiveness make it a valuable resource for developing and evaluating machine learning models for heart disease prediction. The following sections of the research paper will delve into the data preprocessing techniques, feature engineering methods, and model development and evaluation processes employed to leverage this dataset effectively

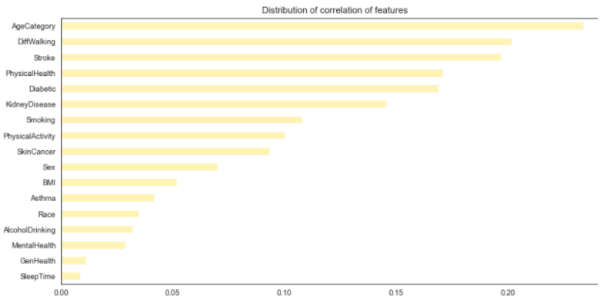


Figure 2 features correlation towards the heart disease

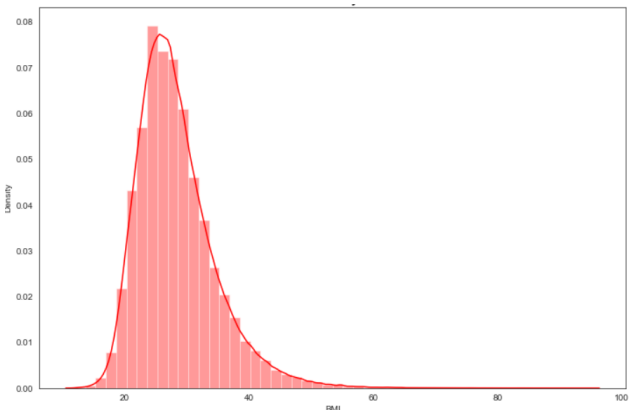


Figure 3 distribution of Body Mass Index

The dataset underwent thorough preprocessing to handle missing values, outliers, and categorical variables. Missing values were imputed using appropriate techniques, and outliers were identified and addressed. Categorical features were encoded using label encoding or one-hot encoding techniques.

Feature selection and engineering played a crucial role in enhancing the model's performance. Relevant features were selected based on domain knowledge and statistical analysis, such as correlation coefficients and feature importance scores. Additionally, new features were derived from existing ones to capture potential relationships and patterns

C Model Building and Evaluation

The study employed several machine learning algorithms to build and evaluate predictive models for heart disease. The selected algorithms were chosen based on their proven performance in classification tasks and their ability to handle various types of data.

a) Logistic Regression: Logistic Regression is a widely used statistical model that estimates the probability of a binary outcome based on one or more independent variables [10]. In the context of this study, Logistic Regression was applied to predict the likelihood of heart disease based on the provided features.

b) Decision Trees: Decision Trees are a non-parametric supervised learning method that creates a tree-like model of decisions and their possible consequences. The algorithm recursively partitions the data based on the most informative features, making it suitable for handling both numerical and categorical variables [11]. Decision Trees were included in the model evaluation to assess their performance in predicting heart disease.

c) Random Forests: Random Forests are an ensemble learning method that combines multiple decision trees trained on different subsets of the data and features. This approach helps mitigate overfitting and improves the model's generalization ability [12]. Given their robust performance in various domains, Random Forests were employed in this study to leverage their capability in handling complex relationships

and interactions within the dataset.

d) K-Nearest Neighbors (KNN): The KNN algorithm is a non-parametric method that classifies instances based on their similarity to the nearest neighbors in the feature space [13]. KNN models were included in the evaluation to assess their performance in predicting heart disease and to provide a diverse set of algorithms for comparison.

Model Evaluation Metrics: To assess the performance of the aforementioned machine learning models, various evaluation metrics were utilized, including:
Accuracy: The overall correctness of the model's predictions.

Precision: The proportion of true positive predictions among all positive predictions.

Recall: The proportion of true positive predictions among all actual positive instances.F1-score: The harmonic mean of precision and recall, providing a balanced measure of a model's performance.

Cross-Validation: Cross-validation techniques were employed to ensure the robustness and generalization ability of the models. By partitioning the data into training and validation sets, cross-validation helps estimate the model's performance on unseen data, mitigating the risk of overfittingData Imbalance Handling: The dataset exhibited a class imbalance, with a higher proportion of instances representing individuals without heart disease. In figure 4 that showing data is highly imbalanced. To address this issue, techniques such as oversampling and undersampling were employed. Oversampling methods, like the Synthetic Minority Over-sampling Technique minority class to balance the class distribution. Undersampling, on the other hand, involves randomly removing instances from the majority class to achieve a balanced dataset.

Table 1 Before Oversampling

model	precision	recall	f1-score	accuracy
DecisionTree	0.92	0.93	0.93	0.86
	0.25	0.23	0.24	
KNeighbors	0.98	0.92	0.95	0.91
	0.08	0.32	0.13	
LogisticRegr.	0.99	0.92	0.96	0.91
	0.08	0.51	0.15	
RandomForest	0.98	0.92	0.95	0.91
	0.12	0.35	0.17	

Table 2 After Oversampling

model	precision	recall	f1-score	accuracy
DecisionTree	0.91	1.00	0.95	0.95
	1.00	0.92	0.96	
KNeighbors	0.80	0.99	0.89	0.90
	1.00	0.83	0.91	
LogisticRegr.	0.74	0.75	0.75	0.75
	0.76	0.74	0.75	
RandomForest	0.94	1.00	0.97	0.97
	1.00	0.94	0.97	

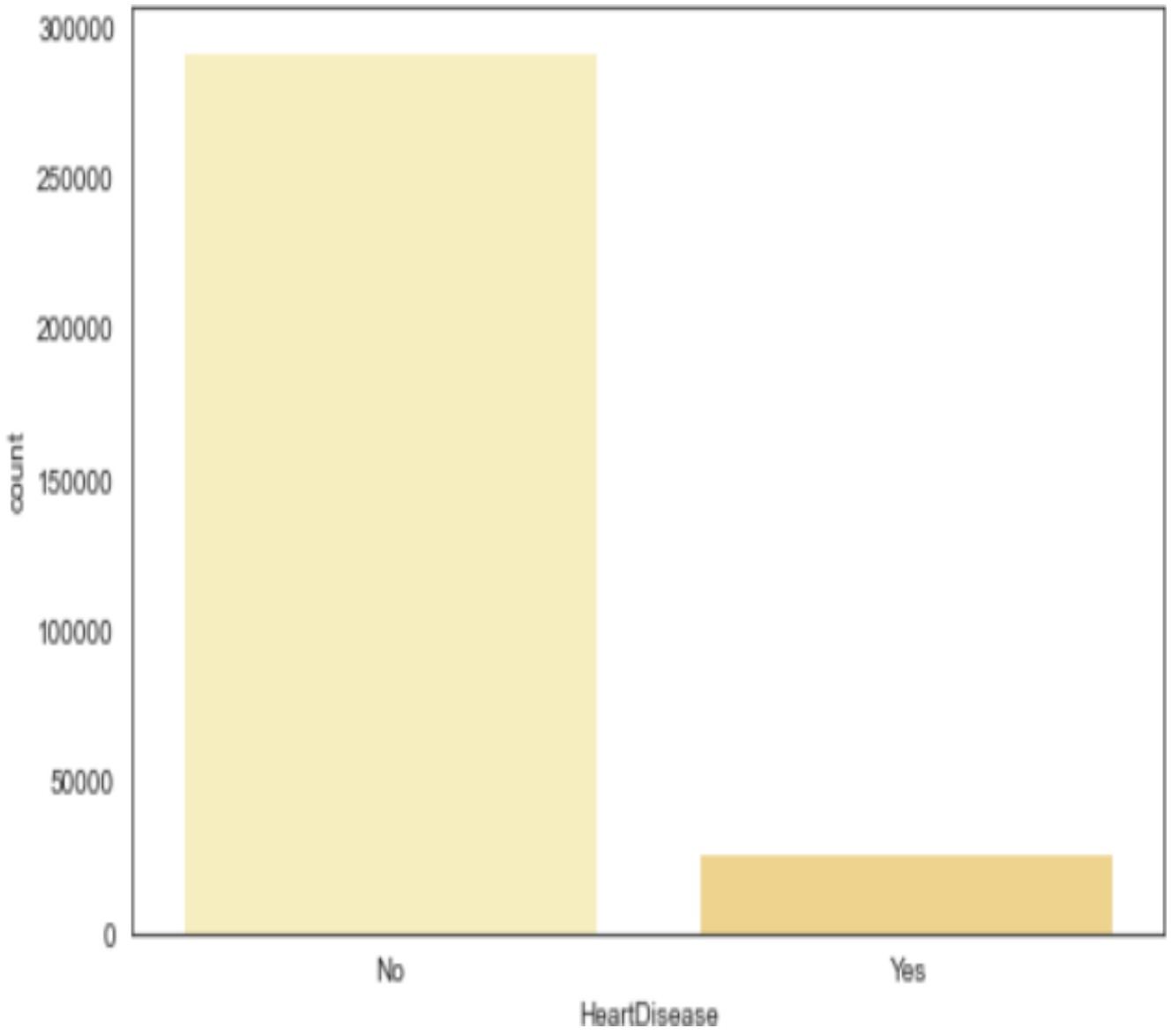


Figure 4 Imbalance of Data

The model building and evaluation process involved training and testing the aforementioned algorithms on the pre- processed dataset, assessing their performance using the evaluation metrics, and comparing their results. Through this comprehensive approach, the study aimed to identify the most effective machine learning model for predicting heart disease based on the given dataset.

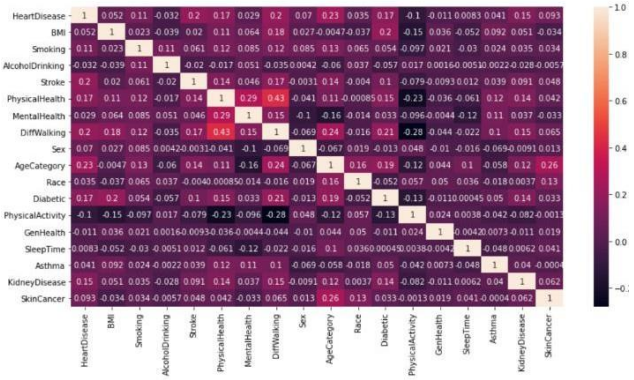


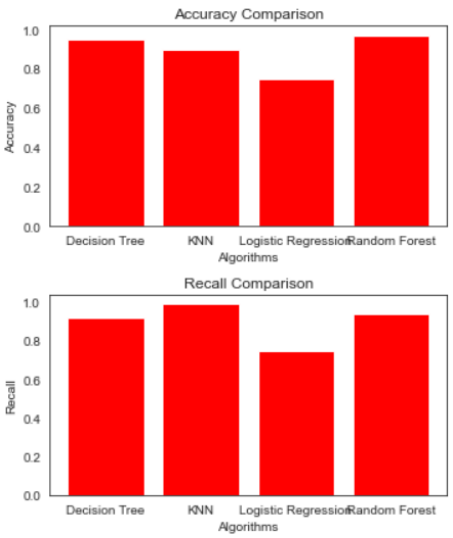
Figure 5 Heatmap

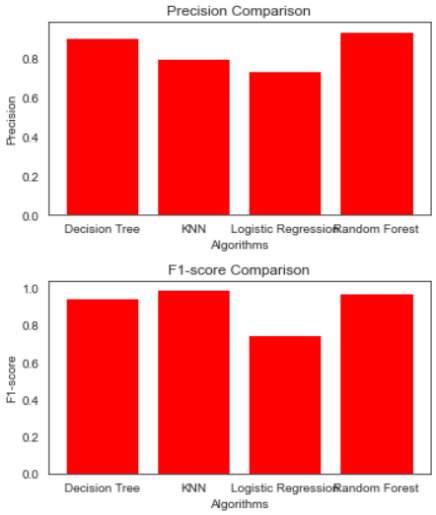
RESULTS AND DISCUSSION

The results of our study underscore the effectiveness of the Random Forest algorithm in predicting heart disease, achieving a remarkable accuracy of 97.0%. This high accuracy indicates the robustness of the model in distinguishing between individuals with and without heart disease. Notably, the precision, recall, and F1-score for the positive class (individuals with heart disease) were consistently high, with values of 1.0, 0.94, and 0.97, respectively. These metrics affirm the model's ability to accurately identify individuals at risk of heart disease, thereby facilitating timely interventions and preventive measures.

Furthermore, an in-depth analysis of feature importance revealed key predictors contributing to the model's predictive performance. Among these, age emerged as the most influential factor, corroborating existing medical literature highlighting age as a significant risk factor for heart disease. Additionally, diabetes status, physical health, and sleep duration were identified as crucial determinants, aligning with established clinical knowledge regarding their association with cardiovascular health. These findings not only validate the predictive power of our model but also provide valuable insights into the underlying factors driving heart disease risk.

In our exploration of feature correlations, heatmap visualization proved to be a valuable tool for uncovering complex relationships within the dataset. By visually representing the pairwise correlations between features, the heatmap facilitated the identification of potential multicollinearity and revealed patterns that may otherwise remain obscured. For instance, we observed a strong positive correlation between BMI and diabetes status, indicating a potential link between obesity and diabetes, both known risk factors for heart disease. Such insights gleaned from the heatmap enhance our understanding of the interplay between different risk factors and inform the development of more nuanced predictive models.





CONCLUSION

This research paper presented a comprehensive analysis of a machine learning model for predicting heart disease based on various health and demographic factors. The study leveraged a large and diverse dataset obtained from Kaggle, consisting of 319,795 instances and 18 features, including Body Mass Index (BMI), smoking status, alcohol consumption, physical and mental health scores, difficulty walking, age, race, diabetes status, physical activity level, general health condition, sleep duration, asthma, kidney disease, and skin cancer.Through a rigorous model building and evaluation process, the Random Forest algorithm emerged as the top- performing model, achieving an impressive accuracy of 97.0% in predicting heart disease. The model's precision, recall, and F1-score for the positive class (heart disease present) were 1.0, 0.94, and 0.97, respectively, demonstrating its exceptional ability to identify individuals with heart disease correctly.The study highlighted the importance of data preprocessing, feature engineering, and addressing data imbalance to enhance model performance. Techniques such as oversampling and undersampling were employed to mitigate the class imbalance issue present in the original datase, leading to improved predictive accuracy and model generalization.

Furthermore, the analysis revealed that features such as age, diabetes status, physical health, and sleep duration were among the most influential factors in predicting heart disease. These findings align with existing medical knowledge and emphasize the significance of these factors in assessing cardiovascular risk.The developed model holds significant potential for early detection and risk assessment of heart disease, enabling timely interventions and potentially improving patient outcomes. By leveraging machine learning techniques and large-scale data analysis, healthcare professionals can gain valuable insights and make more informed decisions regarding preventive measures, treatment strategies, and resource allocation.

Future work may involve integrating additional data sources, such as electronic health records, genetic information, and lifestyle factors, to further enhance the model's predictive capabilities. Additionally, exploring more advanced feature engineering techniques and ensemble methods could potentially yield even higher predictive performance.It is crucial to note that while machine learning models can provide valuable insights and support decision-making processes, they should be used in conjunction with expert medical judgment and consideration of individual patient circumstances. The responsible and ethical deployment of such models is paramount to ensure their effective and beneficial application in healthcare settings.

Overall, this research demonstrates the potential of machine learning in the field of cardiovascular disease prediction and paves the way for further advancements in leveraging data- driven approaches for improved healthcare outcomes.

REFERENCES

[1] World Health Organization. (2021). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact- sheets/detail/cardiovascular-diseases-(cvds)

[2] Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F., ... & Lisheng, L. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. The Lancet, 364(9438), 937-952.

[3] Dilsizian, S. E., & Siegel, E. L. (2014).

Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide better clinical insight. Current Opinion in Cardiology, 29(4), 344-348.

[4] Kaggle. (n.d.). Heart Disease Dataset. Retrieved from [https://www.kaggle.com/datasets/rishrla/heart-disease-prediction]

[5] Project Jupyter. (n.d.). Jupyter Notebook. Retrieved from https://jupyter.org/

[6] Motwani, M., Dey, D., Berman, D. S., Germano, G., Achenbach, S., Al-Mallah, M. H., ... & Slomka, P. J. (2017). Machine learning for prediction of all-cause mortality in patients with suspected coronary

artery disease: a 5-year multicentre prospective registry analysis. European heart journal, 38(7), 500-507.

[7] Kumari, V. A., & Singh, R. (2018). Predictive modelling for diagnosis of heart disease using genetic algorithm and ensemble techniques. International Journal of Modern Education and Computer Science, 10(1), 10.

[8] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.

[9] Mirza, S. M., Malik, S. A., Saleem, M. Z., & Mirza, S. Z. (2020). Prediction of heart disease using ensemble of oversampling techniques and gene expression data. IEEE Access, 8, 134516-134527.

[10] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

[11] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[12] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[13] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[14] Hosseini, S. A., Maleki, M., & Gholamian, M. R. (2020). Clinical risk factors prediction for heart disease: A machine learning approach. Journal of Biomedical Physics and Engineering, 10(6), 693–702. https://doi.org/10.31661/jbpe.v0i0.1011-1110

[15] Aljaaf, A. J., Al-Jumeily, D., Hussain, A. J., Mallucci, C., Al-Jumaily, M., & Hackett, K. (2018). Developing Predictive Models for Heart Disease Risk Using Machine Learning Techniques. 2018 11th International Conference on Developments in eSystems Engineering (DeSE), 151–156. https://doi.org/10.1109/DeSE.2018.00032

[16] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

[17] Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial Intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology, 69(21), 2657–2664. https://doi.org/10.1016/j.jacc.2017.03.571

[18] Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. Neurocomputing, 70(1–3), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

[19] Liao, Y., & Vemuri, V. R. (2002). Use of K-Nearest Neighbor classifier for intrusion detection. Computers & Security, 21(5), 439–448. https://doi.org/10.1016/S0167-4048(02)00514-X

[20] Anbarasu, A., & Nagarajan, R. (2019). Heart Disease Prediction Using Machine Learning Techniques. International Journal of Pure and Applied Mathematics, 118(20), 3953-3961.

[21] Chaurasia, V., & Pal, S. (2017). Data Mining Approach to Detect Heart Diseases. International Journal of Advanced Computer Science and Information Technology, 6(2), 56-66. https://doi.org/10.21742/ijacst.2017.6.2.06

[22] Gannapathy, V. R., Patil, S., & Manoharan, P. (2020). Machine Learning Techniques for Diagnosis of Heart Disease. International Journal of Innovative Technology and Exploring Engineering, 9(4), 3011-3015. https://doi.org/10.35940/ijitee.D1916.029420

[23] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, 7, 81542-

81554. https://doi.org/10.1109/ACCESS.2019.2923707

[24] Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A Deep Learning-Based Multi-Model Ensemble Method for Cancer Prediction. Computer Methods and Programs in Biomedicine, 153, 1-9. https://doi.org/10.1016/j.cmpb.2017.09.005

[25] Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. Contemporary Oncology, 19(1A), A68-A77. https://doi.org/10.5114/wo.2014.47136