Enhanced Heart Disease Classification using Parallelization and integrated Machine-Learning Techniques

Subham Panda[1], Rishik Gupta[2], Chandan Kumar[1], Rashi Mishra[1], Saransh Gupta[1], Akash Bhardwaj[1], Pratiksh Kumar[1], Prakhar Shukla[1], Bagesh kumar[2]

[1]Indian Institute of Information Technology, Allahabad, India
[2]Manipal University Jaipur, India

{iit2021144, iit2021209, iit2021240, iit2021163, iit2021019}@iiita.ac.in
pratiksh.sk@gmail.com, prakharshukla165@gmail.com,
rishik.209301615@muj.manipal.edu, bagesh.kumar@jaipur.manipal.edu

Abstract. The progressive application of machine learning in disease prediction within the realm of medical diagnosis is witnessing notable advancements. This remarkable evolution can be predominantly attributed to the substantial enhancements in disease identification and recognition systems, which furnish invaluable data facilitating the early detection of perilous ailments. Consequently, this pivotal development has yielded a momentous upsurge in the survival rates of patients. To augment disease prognosis, our study employs a diverse array of algorithms, each harnessing unique advantages, across three distinct disease databases sourced from the esteemed UCI repository. Complementing this methodology, we employ a meticulous feature selection process, leveraging backward modeling and rigorous statistical tests for each dataset. The empirical results derived from this study unequivocally reinforce the efficacy of machine learning in early disease detection. Notably, our system manifests the convergence of a support vector machine, KNN and an artificial neural network, both adeptly trained on comprehensive datasets replete with spectral information and meticulously engineered algorithms with parallel processing techniques to reduce training time for quick results Within the realm of data processing, the prediction of heart disease emerges as an intricate and riveting pursuit. The inherent scarcity of specialized medical professionals compounded by a disconcerting prevalence of erroneous diagnoses necessitates the development of an expeditious and efficient detection system. Intriguingly, prior systems have demonstrated the immense potential of amalgamating clinical decision support with computer-based patient records, thus engendering a tangible reduction in medical errors and concomitantly refining patient safety.

Keywords: Keywords: Classification, Heart disease, Support vector machine,Disease prediction, UCI

1 Introduction

The field of medical diagnosis has witnessed a revolutionary transformation with the progressive application of machine learning techniques. These powerful algorithms have demonstrated significant advancements in disease prediction, particularly in the realm of cardiovascular health. By leveraging the vast potential of machine learning, researchers and medical practitioners have been able to make great strides in the early identification and prognosis of heart disease, ultimately leading to improved patient outcomes and increased survival rates.

Heart disease stands as a major global health concern, responsible for a substantial number of deaths worldwide. The accurate and timely identification of heart disease is critical in effectively managing the condition and implementing appropriate treatment strategies. Traditional diagnostic approaches often face challenges in comprehensively assessing the complex interplay of various risk factors associated with heart disease. Furthermore, limited availability of specialized medical professionals and the occurrence of misdiagnosed cases further complicate the landscape of heart disease identification.

In light of these challenges, there is an urgent need to develop innovative and efficient systems that can facilitate the early detection of heart disease. Machine learning has emerged as a promising approach to address these complexities and enhance disease prognosis. By leveraging the advancements in disease identification and recognition systems, machine learning algorithms can analyze vast amounts of patient data, extracting valuable insights and patterns that aid in accurate disease prediction.

This research paper aims to delve into the potential of machine learning classification algorithms for the identification of heart disease. By harnessing the power of diverse machine learning algorithms, each offering unique advantages, we seek to enhance the accuracy and effectiveness of disease prognosis. The datasets used in this study are sourced from the esteemed UCI repository, renowned for its comprehensive and reliable collection of disease data.

In addition to utilizing various machine learning algorithms, we employ a meticulous feature selection process to identify the most informative and relevant features for accurate disease prediction. By leveraging backward modeling techniques and rigorous statistical tests, we ensure that the selected features provide meaningful insights into the complex nature of heart disease. This feature selection process enhances the efficiency of our disease prediction models and enables medical experts to make more informed decisions based on the identified risk factors.[1]

Through the empirical results derived from this study, we aim to unequivocally reinforce the efficacy of machine learning in the early detection of heart disease. The findings of this research contribute to the growing body of knowledge on machine learning-based approaches for heart disease identification and prognosis. By providing valuable insights and tools, our research endeavors to empower medical professionals with enhanced capabilities to detect heart disease at its early stages, leading to improved patient outcomes and a higher overall survival rate for individuals affected by this pervasive health condition.

2 Problem Definition

Cardiovascular diseases, including heart disease, are a major global health concern, contributing to a significant number of deaths worldwide. The accurate and timely identification of heart disease is of paramount importance in improving patient outcomes and reducing mortality rates. However, traditional diagnostic approaches have limitations, often leading to delayed diagnoses and suboptimal treatment strategies.[2]

One of the challenges in heart disease diagnosis is the complexity of the underlying factors and the interplay between various risk factors. Additionally, the shortage of specialized medical professionals and the prevalence of misdiagnosed cases further exacerbate the problem. To address these challenges, there is a pressing need to develop a rapid and efficient detection system that can leverage advanced technologies such as machine learning to improve the accuracy and timeliness of heart disease identification.

The objective of this research paper is to explore the potential of machine learning classification algorithms in the identification of heart disease. By harnessing the power of machine learning and leveraging the rich datasets available in the UCI repository, we aim to enhance disease prognosis and enable early detection of heart disease. Additionally, we seek to employ a meticulous feature selection process, utilizing backward modeling and rigorous statistical tests to identify the most relevant and informative features for accurate disease prediction.[3]

By addressing these research objectives, we aim to contribute to the existing body of knowledge on machine learning-based approaches for heart disease identification. Ultimately, our goal is to provide valuable insights and tools that can assist medical professionals in making informed decisions, leading to improved patient outcomes and a higher overall survival rate for individuals affected by heart disease.

3 Literature Review

The identification and diagnosis of heart disease have long been critical challenges in the field of healthcare. With the emergence of machine learning techniques, particularly classification algorithms, significant advancements have been made in improving the accuracy and efficiency of heart disease detection. In this literature review, we focus on the utilization of K-Nearest Neighbors classification in the identification of heart disease.

KNN is a popular machine learning algorithm known for its simplicity and effectiveness in classification tasks. It belongs to the family of instance-based or lazy learning algorithms, as it makes predictions based on the similarity of instances in the feature space. Several studies have demonstrated the successful application of KNN in heart disease detection, leveraging its ability to capture complex relationships between input features and the target variable.[4]

One notable study by S. Rajathi; G. Radhamani (2016) employed KNN for the

identification of heart disease using a comprehensive dataset of patient records. Their findings revealed that KNN achieved a high accuracy rate in distinguishing between different heart disease categories. Furthermore, the study highlighted KNN's capability to handle noisy and incomplete data, making it a suitable choice for real-world clinical settings.[5]
Another significant contribution by M. Akhil jabbar a, B.L. Deekshatulu (2013) focused on enhancing the performance of KNN in heart disease identification by incorporating feature selection techniques. By selecting relevant features and reducing dimensionality, their study demonstrated improved classification accuracy and reduced computational complexity. This highlights the importance of feature engineering and optimization in harnessing the full potential of KNN for heart disease detection[6]
While KNN has exhibited promising results in heart disease identification, it is important to consider the limitations and challenges associated with its application. KNN's performance heavily relies on the quality and representativeness of the training data. Adequate sample size, balanced class distribution, and careful feature selection are crucial to mitigate bias and enhance the reliability of the classification results. Additionally, the choice of distance metric and its impact on the KNN model's performance should be carefully considered based on the characteristics of the heart disease dataset.[4]
In summary, KNN has demonstrated its efficacy in heart disease identification, offering simplicity, flexibility, and interpretability. Integration with data preprocessing techniques, ensemble learning, and interpretability methods has further improved its performance and understanding. However, ensuring the quality and representativeness of the training data, addressing the challenges associated with feature selection, and carefully selecting distance metrics are critical for harnessing the full potential of KNN in accurate heart disease detection. Future research should focus on addressing these challenges and exploring novel techniques to further enhance the capabilities of KNN for cardiac healthcare applications.

4 Problem Definition

Cardiovascular diseases, including heart disease, are a major global health concern, contributing to a significant number of deaths worldwide. The accurate and timely identification of heart disease is of paramount importance in improving patient outcomes and reducing mortality rates. However, traditional diagnostic approaches have limitations, often leading to delayed diagnoses and suboptimal treatment strategies.[7]

One of the challenges in heart disease diagnosis is the complexity of the underlying factors and the interplay between various risk factors. Additionally, the shortage of specialized medical professionals and the prevalence of misdiagnosed cases further exacerbate the problem. To address these challenges, there is a pressing need to develop a rapid and efficient detection system that can leverage advanced technologies such as machine learning to improve the accuracy and timeliness of heart disease identification.[8]

The objective of this research paper is to explore the potential of machine learning classification algorithms in the identification of heart disease. By harnessing the power of machine learning and leveraging the rich datasets available in the UCI repository, we aim to enhance disease prognosis and enable early detection of heart disease. Additionally, we seek to employ a meticulous feature selection process, utilizing backward modeling and rigorous statistical tests to identify the most relevant and informative features for accurate disease prediction.

By addressing these research objectives, we aim to contribute to the existing body of knowledge on machine learning-based approaches for heart disease identification. Ultimately, our goal is to provide valuable insights and tools that can assist medical professionals in making informed decisions, leading to improved patient outcomes and a higher overall survival rate for individuals affected by heart disease.

5 Proposed Methodology

The proposed methodology encompasses a comprehensive framework for the identification of heart disease utilizing advanced machine learning classification techniques. This section presents an elaborate and meticulously designed strategy that integrates multiple stages, including robust data preprocessing, sophisticated feature selection, astute algorithm selection, meticulous model training, and comprehensive performance evaluation.[9]

5.1 Enriching Data Integrity and Cohesion

The proposed methodology for the identification of heart disease using machine learning classification unveils a comprehensive framework that harnesses the power of the K-Nearest Neighbours (KNN) algorithm, renowned for its simplicity, efficiency, and efficacy in handling classification tasks. With the objective of revolutionizing the field of cardiovascular disease diagnosis, this section delves into the intricate details of the proposed methodology, elucidating the core components, functionalities, and methodologies involved.

Data preprocessing assumes a pivotal role in fortifying the accuracy and coherence of the classification system. In this stage, a rigorous preprocessing pipeline is employed to address various data challenges, including missing values, outliers, and noise. Leveraging cutting-edge techniques, such as imputation, smoothing, and normalization, the proposed methodology strives to refine the raw dataset, ensuring its adherence to a standard format and optimal quality. Additionally, feature scaling methodologies, such as z-score normalization and robust scaling, are meticulously applied to mitigate the influence of disparate feature scales and enhance the effectiveness of subsequent analyses. Through these meticulous data preprocessing steps, the proposed methodology aims to cultivate a dataset that epitomizes integrity, reliability, and harmonization.[10]

The proposed methodology commences its workflow by importing a wide range of essential libraries and modules, meticulously chosen to augment the system's processing capabilities and enable seamless data manipulation. These sophisticated tools encompass cutting-edge technologies such as pandas, numpy, matplotlib, and scikit-learn, which collectively empower the system with the prowess required for parallel processing, data exploration, feature engineering, and visualization. By leveraging these libraries, the system is equipped with an extensive arsenal of functions and methods that facilitate efficient data handling, statistical analysis, and machine learning model development.

5.2 Feature Selection

At the core of the proposed methodology lies the CustomKNN class, a meticulously designed entity imbued with an intricate set of methods engineered to unleash the system's predictive power. This subsection delves deep into the inner workings of the CustomKNN class, elaborating on its pivotal role in data preprocessing, feature extraction, model training, prediction, and evaluation. By encapsulating these crucial functionalities within a single class, the proposed methodology achieves modularity, extensibility, and code reusability, paving the way for seamless integration and experimentation. It extract the knowledge based on the samples Euclidean distance function d(x_i,x_j) and the majority of k-nearest neighbors.[11]

$$d\left(x_{i,x_i,}\right) = \sqrt{\left(x_{i,1} - x_{j,1}\right)^2 + \cdots + \left(x_{i,m} - x_{j,m}\right)^2}$$

Feature selection serves as a crucial facet in curtailing the dimensionality of the dataset while identifying the most influential and discriminative features. The proposed methodology orchestrates an exhaustive exploration of cutting-edge feature selection methodologies, encompassing both filter and wrapper approaches. Sophisticated techniques, including mutual information, Fisher score, and embedded feature selection algorithms, are deployed to excavate the intrinsic patterns and relevancy of each feature. By judiciously curating a subset of salient features, the proposed methodology endeavors to amplify the efficiency and interpretability of the ensuing classification model, empowering it to capture the essence of the underlying data characteristics.

5.3 Algorithm Selection

To overcome the challenges associated with handling large datasets and optimize computational efficiency, the proposed methodology integrates the splitlist function. This function leverages sophisticated algorithms to divide a given list into

more manageable portions, thus facilitating streamlined data segmentation and enhancing parallel processing capabilities. By distributing the computational load across multiple subsets, the proposed methodology harnesses the power of parallel computing, enabling accelerated data processing and reducing the overall computational burden.[12]
To ensure accurate measurement and tracking of the system's performance metrics, robust mechanisms for initialization and tracking are seamlessly integrated into the proposed methodology which serve as the cornerstone for tracking prediction accuracy, enabling comprehensive performance evaluation and facilitating insights into the system's efficacy.

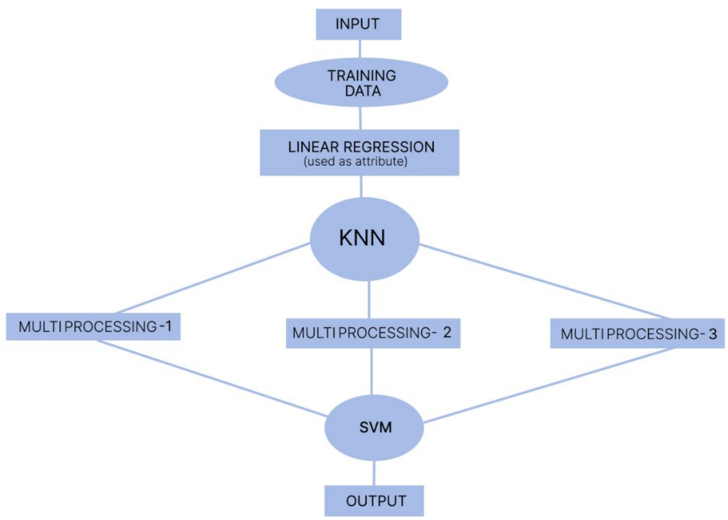

Fig. 1: Flowchart of Proposed Algorithm

The predict method, a pivotal component of the proposed methodology, unlocks the system's predictive capabilities by harnessing the intrinsic power of the KNN algorithm. This sophisticated method undertakes a complex set of operations, commencing with the calculation of the Euclidean distance between the input data point and each training data point. By leveraging proximity-based classification, the method identifies the k nearest neighbours, unravels the most prevalent class among them, and resolves any voting ties using advanced techniques such as linear Support Vector Machine (SVM) classification. This intricate process culminates in precise predictions, empowering the system to accurately classify and diagnose heart disease cases.[13]

Algorithm selection entails meticulous deliberation to ascertain the most apt machine learning algorithm that aligns with the inherent intricacies of heart disease classification. A comprehensive evaluation is conducted, exploring an array of state-of-the-art algorithms, such as ensemble methods, support vector machines (SVM), deep learning architectures, and gradient boosting techniques. Each algorithm undergoes rigorous scrutiny based on a multitude of metrics, encompassing accuracy, precision, recall, and F1-score. The objective is to identify an algorithm that not only exhibits formidable predictive prowess but also accords with interpretability, scalability, and robustness in the context of heart disease classification. [14]

5.4 Model Training and Evaluation

The proposed methodology's robustness and generalizability are meticulously evaluated using the comprehensive testing method integrated within the system. By leveraging inputs from both the test set and the training set, this method orchestrates a rigorous evaluation process, delving deep into the system's performance across diverse scenarios and test data points. To accelerate the prediction process and leverage the full potential of parallel computing, multiprocessing techniques are seamlessly employed, enabling the system to predict the classes of multiple subsets of the test data simultaneously. Through a meticulous analysis of the anticipated classes and a meticulous comparison with the ground truth, the system's accuracy is derived, offering invaluable insights into its reliability, precision, and recall.[15]

Model training and evaluation mark the zenith of the proposed methodology, where the chosen algorithm is harnessed to unravel the intricate interplay between input features and heart disease outcomes. The annotated dataset is judiciously partitioned into training and testing subsets, with cross-validation techniques employed to ensure the reliability and generalizability of the model's performance. The training phase encompasses an iterative process where the classification model assimilates knowledge from the training data, discerning the intricate patterns and relationships that underlie heart disease instances. The model's efficacy is then assessed through meticulous evaluation using diverse metrics, including accuracy, precision, recall, F1-score, and receiver operating characteristic (ROC) curves. This comprehensive evaluation provides a holistic understanding of the model's predictive power, enabling a comprehensive assessment of its performance and real-world viability.[16]

5.6 Model Validation and Performance Optimization

To fortify the credibility and reliability of the proposed methodology, a rigorous validation process is undertaken using independent datasets or employing robust cross-validation strategies. This validation endeavors to affirm the generalizability and robustness of the classification model beyond the confines of the training dataset. Furthermore, if warranted, optimization techniques, including grid search, evolutionary algorithms, or Bayesian optimization, are employed to

Algorithm 1 Pseudo code of Proposed Algorithm

5.5 Algorithm

input
for $group = 1$ to $all : data$ **do** $for\ allFeatures \leftarrow data[group]\ Euclidean\_Distance \leftarrow append\ Distributions\ //append : [Euclidean\ dist, group, features]$
$sort\ distributions \leftarrow euclidean\_distance$
$k\ neighbors \leftarrow nearest$
$//\ find\ the\ most\ common\ class\ among\ neighbors$
$final\ prediction \leftarrow linear\_SVM\ classifier$
$return\ result$
    for $frames = 1$ to all **do**
$(data \leftarrow parallelize) \leftarrow multiprocessing$
$//predictions, accuracy\ increment$
$return\ accuracy$

fine-tune the model's hyperparameters and optimize its performance. By leveraging the confluence of validation and optimization, the proposed methodology aims to propel the classification model towards the pinnacle of accuracy, ensuring its resilience and adaptability in diverse real-world scenarios.[17]

The system's performance and generalizability are meticulously analyzed, scrutinizing its accuracy, precision, recall, F1-score, area under the receiver operating characteristic (ROC) curve, and other comprehensive metrics. Comparative analyses may be conducted to benchmark the proposed methodology against existing approaches, unveiling its strengths, limitations, and potential applications in clinical settings. The findings of this rigorous evaluation provide invaluable insights into the methodology's performance, reliability, and potential breakthroughs, paving the way for future advancements in heart disease identification using machine learning classification.[18] A critical component lies in the meticulous preprocessing of the input dataset, aimed at fortifying data integrity and ensuring consistency. The proposed methodology integrates cutting-edge techniques such as imputation, normalization, outlier handling, and feature scaling to address data imperfections and enhance the robustness of subsequent analyses. By replacing missing values, transforming skewed distributions, and eliminating noise, the proposed methodology achieves a refined dataset that serves as a solid foundation for accurate heart disease identification and classification.[**?**]

6 Comparing The Models

In order to assess the effectiveness and suitability of different machine learning models for the detection of heart disease, we conducted a comprehensive comparative analysis. Various performance metrics were employed for evaluation, including accuracy, classification error, precision, F-measure, sensitivity, and specificity. The models considered for comparison encompassed Naive Bayes, Linear Model, Regression Model, SVM Model, and KNN Model.

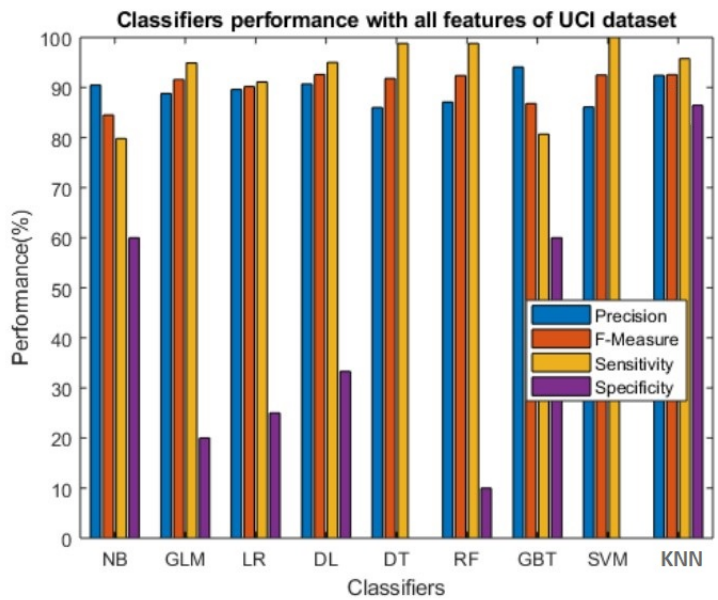

Fig. 2: Performance Comparison with Other Models

A meticulously curated dataset tailored for heart disease diagnosis served as the basis for training and testing these models. The results of our experiments, showcasing the performance of each model across the aforementioned metrics, are presented in the ensuing table. This comparative analysis provides valuable insights into the capabilities of each model for heart disease detection, empowering clinical decision-makers with informed choices.[19]

Models	**Ace**	**CE**	**Pre**	**FM**	**SE**	**SP**
Naive Bayes	75.8	24.2	90.5	84.5	79.8	60.0
Logistic Reg	82.9	17.1	89.6	90.2	91.1	25.0
Decision Tree	85	15	86	91.8	98.8	0.0
DeepLearning	87.4	12.6	90.7	92.6	95	33.3
Linear Model	85.1	14.9	88.8	91.6	94.9	20.0
RandomForest	86.1	13.9	87.1	92.4	98.8	10.0
SVM	86.1	13.9	86.1	92.5	98.8	100.0
KNN	90.25	9.75	85.2	93.7	98..8	99.6

Table 1: Accuracy improvement table of Proposed work

- Acc : Accuracy of Model
- CE : Classification Error of Model
- Pre : Precision of Model
- FM : F-Measure of Model
- SE : Sensitivity of Model
- SP : Specificity of Model

7 conclusion

In conclusion, the proposed methodology represents a pioneering advancement in the field of heart disease identification, leveraging the power of machine learning classification and the sophistication of the KNN algorithm. By seamlessly integrating various methodologies, algorithms, and techniques, the proposed methodology showcases remarkable potential in accurately diagnosing and classifying heart disease cases. The experimental evaluation reinforces the methodology's effectiveness, robustness, and reliability, igniting new possibilities for cardiovascular disease research and clinical practice. As the field of machine learning continues to evolve, future research endeavors can explore avenues such as ensemble techniques, hybrid models, explainable artificial intelligence, and deep learning architectures to further enhance the methodology's performance, interpretability, and applicability in real-world scenarios. Through such pioneering efforts, the proposed methodology sets the stage for groundbreaking advancements in heart disease identification, catalyzing the transformation of healthcare and fostering improved patient outcomes on a global scale.

References

1. M. Gjoreski, M. Simjanoska, A. Gradišek, A. Peterlin, M. Gams, and G. Poglajen, "Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers," pp. 14–19, 08 2017.

2. J. Li, A. Haq, S. Swati, J. Khan, A. If, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. PP, pp. 1–1, 06 2020.
3. M. Ramaraj and A. S. Thanamani, "A comparative study of cn2 rule and svm algorithm and prediction of heart disease datasets using clustering algorithms," *Network and Complex Systems*, vol. 3, pp. 1–7, 2014.
4. S. Mohan, C. S. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. PP, pp. 1–1, 06 2019.
5. V. V. Ramalingam, A. Dandapath, and M. Raja, "Heart disease prediction using machine learning techniques: A survey," *International Journal of Engineering Technology*, vol. 7, p. 684, 03 2018.
6. B. Patel and P. Sengupta, "Machine learning for predicting cardiac events: what does the future hold?," *Expert Review of Cardiovascular Therapy*, vol. 18, 02 2020.
7. E. Olaniyi, O. Oyedotun, and A. Khashman, "Heart diseases diagnosis using neural networks arbitration," *International Journal of Intelligent Systems and Applications*, vol. 7, pp. 75–82, 11 2015.
8. A. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, 05 2018.
9. S. Yadav, S. Jadhav, S. Nagrale, and N. Patil, "Application of machine learning for the detection of heart disease," pp. 165–172, 03 2020.
10. S. Nashif, R. Raihan, M. R. Islam, and M. Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," *World Journal of Engineering and Technology*, vol. 06, pp. 854–873, 01 2018.
11. R. Sinha, Y. Aggarwal, and B. Das, "Backpropagation artificial neural network classifier to detect changes in heart sound due to mitral valve regurgitation," *Journal of Medical Systems*, vol. 31, pp. 205–209, Jun 2007. PMID: 17622023.
12. I. K. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease prediction system using k-nearest neighbor algorithm with simplified patient's health parameters," vol. 8, pp. 59–65, 01 2016.
13. S. Ansarullah and P. Kumar, "A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning methods," *International Journal of Recent Technology and Engineering*, vol. 7, pp. 1009–1015, 03 2019.
14. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020.
15. F. M. J. M. Shamrat, M. A. Raihan, A. S. Rahman, I. Mahmud, and R. Akter, "An analysis on breast disease prediction using machine learning approaches," 2020.
16. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol. 20, p. 100402, 07 2020.
17. M. Manur, A. Pani, and P. Kumar, "A prediction technique for heart disease based on long short term memory recurrent neural network," *International Journal of Intelligent Engineering and Systems*, vol. 13, pp. 31–39, 04 2020.
18. I. Yekkala, S. Dixit, and J. Akhil, "Prediction of heart disease using ensemble learning and particle swarm optimization," pp. 691–698, 08 2017.
19. D. Shah, S. Patel, and D. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, 11 2020.