CHAPTER 2
LITERATURE SURVEY

2.1 INTRODUCTION

The literature review section of this project serves as a comprehensive exploration of existing research and studies related to heart disease prediction. It delves into a vast body of literature encompassing epidemiological studies, clinical trials, and machine learning applications in cardiovascular risk assessment. By synthesizing and analyzing previous findings, the literature review provides a solid foundation for understanding the current state of knowledge, identifying gaps, and informing the methodology and approach adopted in this project [40].

2.2 MACHINE LEARNING APPROACHES FOR HEART DISEASE PREDICTION: A REVIEW

In 2024, Amrit Singh et al. proposed a machine learning model for heart disease detection. The proposed framework utilized various machine learning algorithms, including Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Principal Component Analysis (PCA). The model aimed to enhance prediction accuracy and reduce overfitting issues by implementing a novel diagnostic system. The dataset, obtained from the Framingham Heart Study dataset, contained 4,240 records with 16 columns and 15 attributes. The dataset underwent preprocessing, including data cleaning, normalization, and feature selection. Notably, the Random Forest model achieved the highest accuracy of 97%, outperforming other classifiers such as Decision Tree (84%), PCA (79%), and SVM (68%). The study highlights the effectiveness of ensemble learning techniques, particularly Random Forest, in improving heart disease classification accuracy [41].

In 2024, Hosam F. El-Sofany proposed a machine learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Naïve Bayes, Support Vector Machine (SVM), Voting, XGBoost, AdaBoost, Bagging, Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF), and Logistic Regression (LR). The model aimed to enhance prediction accuracy by applying feature selection techniques such as chi-square, analysis of variance (ANOVA), and mutual information (MI). The dataset, obtained from a private hospital dataset and the publicly available Cleveland Heart Disease dataset, contained a total of

503 instances and 13 attributes after preprocessing. The dataset was divided into 75% training and 25% testing sets. Notably, the XGBoost model achieved the highest accuracy of 97.57%, outperforming other classifiers such as Random Forest (93.07%) and AdaBoost (85.15%). The study highlights the effectiveness of ensemble learning techniques, particularly XGBoost, in heart disease classification [42].

In 2023, Chintan M. Bhatt et al. proposed a machine learning model for prediction of cardiovascular disease. The proposed framework utilized various machine learning algorithms, including Decision Tree Classifier (DT), Random Forest (RF), Multilayer Perceptron (MP), and XGBoost (XGB). This model aids in predicting and classifying patients with cardiovascular disease. The dataset, comprised of 70,000 instances, was obtained from the Kaggle. The records were divided into training and testing sets (80:20). Notably, the Multilayer Perceptron model achieved the highest accuracy of 87.28% with cross-validation compared to other ML algorithms [43].

In 2023, Nadikatla Chandrasekhar et al. proposed a machine learning model for prediction of heart disease. The proposed framework utilized various machine learning algorithms. Including Random Forest, K-Nearest Neighbor, Logistic Regression(LR), Naïve Bayes, Gradient Boosting, and AdaBoost(AB) Classifier. This model aids in predicting and classifying patients with heart disease. The dataset comprised of two sources - the Cleveland dataset from the UC Irvine ML Repository (303 instances) and the IEEE Dataport Heart Disease dataset (1190 instances). The results indicated that the LR and AB classifiers attained the highest accuracies of 90.16% and 89.67% on both data sets, respectively. The soft voting ensemble classifier method was applied to all six models on both datasets. It yielded even greater accuracies of 93.44% and 95%. Notably the soft voting ensemble classifier achieved the highest accuracy with cross-validation compared to other ML algorithms [44].

In 2023, Ahmad Ayid Ahmad et al. proposed a machine learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Artificial Neural Networks (ANN), Decision Tree (DT), AdaBoost, and Support Vector Machine (SVM). The model aimed to enhance prediction accuracy by applying the Jellyfish Optimization Algorithm for feature selection. The dataset, obtained from the Cleveland Heart Disease dataset (UCI Repository), contained 1025 instances and 14 attributes. The dataset was divided into 70% training and 30% testing sets. Notably, the SVM model with Jellyfish optimization achieved the highest accuracy of 98.47%, outperforming other classifiers such as

AdaBoost (98.24%), ANN (97.99%), and Decision Tree (97.55%). The study highlights the effectiveness of feature selection using the Jellyfish Optimization Algorithm in improving heart disease prediction accuracy [45].

In 2023, Diaa Salama AbdElminaam et al. proposed a machine learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Logistic Regression, Gradient Boosting, K-Nearest Neighbors (KNN), Random Forest, Naïve Bayes, and Decision Tree. The model aimed to improve heart disease prediction by testing different machine learning classifiers on multiple datasets with varying features. Three datasets were used, containing a total of 574,440 instances with different feature sets (ranging from 12 to 21 attributes). The datasets were divided into 70% training and 30% testing sets. Notably, Logistic Regression achieved the highest accuracy on two datasets (91.6% and 90.8%), while Random Forest achieved the highest accuracy (98.6%) on the third dataset. The study highlights the effectiveness of ensemble learning techniques and dataset variations in enhancing heart disease prediction accuracy [46].

In 2022, Ashish Kumar et al. proposed a deep learning model for early heart attack prediction using Electrocardiogram (ECG) signals. The proposed framework utilized various machine learning algorithms, including Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), Support Vector Machine (SVM), Naïve Bayes, and K-Nearest Neighbors (KNN). The model aimed to improve heart disease prediction by analyzing ECG patterns and optimizing classification techniques. The dataset, obtained from the UCI Machine Learning Repository, contained 383 instances and 14 attributes. The dataset was preprocessed to remove missing values and was divided into training and testing sets. Notably, the CNN model achieved the highest accuracy of 98%, outperforming ANN (94%) and other classifiers. The study highlights the effectiveness of deep learning techniques, particularly CNN, in achieving high accuracy for heart disease prediction based on ECG signals [47].

In 2022, Abdul Saboor et al. proposed a machine learning model for improving the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Random Forest (RF), XGBoost (XGB), Decision Tree (CART), Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), Logistic Regression (LR), Linear Discriminant Analysis (LDA), AdaBoost (AB), and Extra Trees (ET). The model aimed to enhance heart disease prediction through data standardization and hyperparameter tuning. The dataset, obtained from the Cleveland

Heart Disease dataset (UCI Repository), comprised 303 instances with 13 selected attributes. The data was divided into training and testing sets using 10-fold cross-validation. Notably, SVM achieved the highest accuracy of 96.72% after hyperparameter tuning, outperforming other classifiers. The study highlights the effectiveness of data standardization and hyperparameter optimization in improving the accuracy of heart disease prediction models [48].

In 2022, Sashank Yadav et al. proposed a machine learning model for predicting heart disease using 7 machine learning algorithms such as the Naïve Bayes, Decision Tree, Logistic Regression, KNN(K-Nearest Neighbors), SVM(Support Vector Machine), Gradient Boosting and Random Forest algorithms. The heart disease dataset was downloaded from the UCI Machine Learning Repository. The dataset was filtered to include the common 14 features. It consisted of 55% positive and 44.44% negative heart patient data. Among the models KNN achieved a highest accuracy of 85.18% [49].

In 2021, Harshit Jindal et al. proposed a machine learning model for prediction of heart disease. The proposed framework utilized various machine learning algorithms, including logistic regression, K-Nearest Neighbors (KNN), and Random Forest Classifiers. This model aids in predicting and classifying patients with heart disease. The dataset, comprised 13 medical attributes and 304 patient records, was obtained from the UCI Repository. The records were divided into training and testing sets. Notably, the KNN model achieved the highest accuracy of 88.52%, compared to other ML algorithms [50].

In 2021, Dhai Eddine Salhi et al. proposed a machine learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Neural Networks (NN), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The model aimed to improve heart disease prediction by applying feature selection techniques using the Pearson correlation matrix and evaluating different dataset sizes. The dataset, collected from Algerian hospitals, contained 1,200 instances and 13 selected attributes. The dataset was split into 80% training and 20% testing. Notably, the Neural Network model achieved the highest accuracy of 93%, outperforming SVM (90%) and KNN (85.5%). The study highlights the effectiveness of Neural Networks in heart disease prediction, demonstrating its stability across varying dataset sizes [51].

In 2021, Baban U. Rindhe et al. proposed a machine learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest. The model aimed to enhance heart disease prediction accuracy by applying data preprocessing and machine learning classification techniques. The dataset, obtained from the Cleveland Heart Disease dataset (UCI Repository), contained 303 instances and 14 attributes. The dataset was split into training and testing sets. Notably, the Support Vector Classifier achieved the highest accuracy of 84.0%, outperforming Neural Network (83.5%) and Random Forest (80.0%). The study highlights the effectiveness of SVM in heart disease prediction, demonstrating its competitive performance among classification models [52].

In 2021, Apurv Garg et al. proposed a machine learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including K-Nearest Neighbors (KNN) and Random Forest (RF). The model aimed to enhance heart disease prediction accuracy by analyzing key patient attributes such as age, cholesterol levels, and chest pain type. The dataset, obtained from Kaggle (Heart Disease UCI dataset), contained 303 instances and 13 attributes. The dataset was divided into 80% training and 20% testing sets. Notably, the KNN model achieved the highest accuracy of 86.89%, outperforming Random Forest, which achieved 81.97% accuracy. The study highlights the effectiveness of KNN in heart disease prediction, demonstrating its superior performance over Random Forest in this experiment [53].

In 2020, Vijeta Sharma et al. proposed a machine learning model for prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Random Forest, Support Vector Machine (SVM), Naive Bayes, and Decision Tree. This model aids in predicting and classifying patients with heart disease. The dataset, comprised of 1025 instances from the Cleveland Heart Disease Dataset obtained from the UCI Repository, was utilized. The records were divided into training and testing sets. Notably, the Random Forest model achieved the highest accuracy of 99% compared to other ML algorithms [54].

In 2020, Dimas Aryo Anggoro et al. proposed a machine learning model for the prediction of heart disease. The proposed framework compared the performance of two classification algorithms, such as Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). The dataset, consisting of 304 instances and 14 attributes, was obtained from Kaggle, originally compiled from multiple medical institutions. The data was divided into 70% training and 30% testing sets. The study evaluated the impact of normalization on model performance, demonstrating that it significantly improved accuracy. Notably, SVM with normalization achieved the highest accuracy of 90.10%, outperforming KNN, which reached a maximum accuracy of 81.31%. The results suggest that SVM is a more effective algorithm for heart disease prediction, particularly when data normalization is applied [55].

In 2020, Mahesh Parmar et al. proposed a deep learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes, Random Forest, and Deep Neural Networks (DNN). The model aimed to improve heart disease classification using Talos Hyperparameter Optimization. The dataset, obtained from the UCI Heart Disease Repository, comprised 303 instances with 14 selected attributes. The data was divided into training and testing sets, and cross-validation was applied. Notably, the Deep Neural Network optimized using Talos achieved the highest accuracy of 90.78%, outperforming other machine learning classifiers. The study highlights the effectiveness of hyperparameter tuning and deep learning techniques in improving heart disease prediction accuracy [56].

In 2020, R. Jane Preetha Princy et al. proposed a machine learning model for the prediction of cardiac disease. The proposed framework utilized various supervised machine learning algorithms, including Decision Tree, Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The model aimed to enhance disease classification accuracy using a cardiovascular disease dataset from Kaggle, which contained 12 attributes related to patient health conditions. The dataset was divided into 70% training and 30% testing sets. Notably, the Decision Tree model achieved the highest accuracy of 73%, outperforming other algorithms such as Logistic Regression (72%), Random Forest (71%), KNN (66%), and Naïve Bayes (60%). The study highlights the effectiveness of Decision Tree in predicting cardiac disease and suggests that dimensionality reduction can impact model performance [57].

In 2020, Jian Ping Li et al. proposed a machine learning model for the identification of heart disease in e-healthcare. The proposed framework utilized various machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression (LR), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Decision Tree (DT). The model incorporated feature selection techniques, such as Relief, Minimal Redundancy Maximal Relevance (MRMR), Least Absolute Shrinkage and Selection Operator (LASSO), and Local Learning-Based Feature Selection (LLBFS), along with a newly proposed Fast Conditional Mutual Information (FCMIM) feature selection algorithm. The dataset, obtained from the Cleveland Heart Disease dataset (UCI Repository), contained 297 instances and 13 selected attributes after preprocessing. The dataset was splited using the Leave-One-Subject-Out (LOSO) cross-validation method. Notably, the SVM model with the proposed FCMIM feature selection achieved the highest accuracy of 92.37%, outperforming other classifiers. The study highlights the effectiveness of feature selection and machine learning models in improving heart disease prediction accuracy [58].

In 2020, Viren Viraj Shankar et al. proposed a deep learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Convolutional Neural Networks (CNN), Naïve Bayes, and K-Nearest Neighbors (KNN). The model aimed to enhance heart disease prediction accuracy by leveraging structured hospital data, including 13 attributes related to patient health. The dataset was preprocessed to handle missing values and was divided into training and testing sets. Notably, the CNN model achieved the highest accuracy, ranging from 85% to 88%, outperforming the other machine learning classifiers. The study highlights the effectiveness of deep learning, particularly CNN, in improving heart disease prediction accuracy [59].

In 2020, Ilias Tougui et al. proposed a machine learning model for the classification of heart disease. The proposed framework utilized various machine learning algorithms, including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Naïve Bayes, and Random Forest. The study compared the performance of these algorithms across six different data mining tools: Orange, Weka, RapidMiner, Knime, Matlab, and Scikit-Learn. The dataset, obtained from the UCI Machine Learning Repository (Cleveland dataset), contained 303 instances and 13 attributes. The data was processed and evaluated using 10-fold cross-validation, with accuracy, sensitivity, and specificity as performance metrics. Notably, Matlab's Artificial Neural Network model achieved the highest accuracy of 85.86%, outperforming other machine learning classifiers. The study highlights the effectiveness of ANN and Matlab as the best-performing technique and tool for heart disease classification [60].

In 2020, Pooja Anbuselvan et al. proposed a machine learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and XGBoost. The model aimed to enhance prediction accuracy by applying machine learning techniques to medical datasets. The dataset, obtained from the Cleveland Heart Disease dataset (UCI Repository), contained 303 instances and 14 attributes. The dataset was split into 80% training and 20% testing. Notably, the Random Forest model achieved the highest accuracy of 86.89%, outperforming other classifiers such as XGBoost (78.69%) and Decision Tree (77.05%). The study highlights the effectiveness of ensemble learning techniques, particularly Random Forest, in heart disease classification [61].

In 2019, Hager Ahmed et al. proposed a machine learning model for the identification of heart disease from patients' social media posts. The proposed framework utilized various machine learning algorithms, including Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR). The model aimed to improve heart disease prediction by integrating Apache Spark and Apache Kafka for real-time data processing. The dataset, obtained from the Cleveland Heart Disease dataset (UCI Repository), contained 303 instances and 13 attributes. The data was preprocessed using feature selection techniques such as Univariate Feature Selection and Relief and evaluated using k-fold cross-validation. Notably, the Random Forest model achieved the highest accuracy of 94.9%, outperforming other classifiers. The study highlights the effectiveness of real-time machine learning solutions and feature selection in heart disease prediction [62].

In 2019, Abhijeet Jagtap et al. proposed a machine learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB). The model aimed to enhance heart disease prediction accuracy by preprocessing medical datasets and optimizing feature selection. The dataset, obtained from the UCI Machine Learning Repository, was split into 75% training and 25% testing. Notably, the SVM model achieved the highest accuracy of 64.4%, outperforming Logistic Regression (61.45%) and Naïve Bayes (60%). The study highlights the effectiveness of SVM in heart disease classification but suggests the need for further improvements to increase accuracy [63].

In 2017, Sundas Naqeeb Khan et al. proposed a machine learning model for the prediction of heart disease. The proposed framework utilized various machine learning algorithms, including Support Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Tree (C4.5), and RIPPER. The model aimed to compare different classification techniques to determine the most effective method for heart disease prediction. The dataset, obtained from the Cleveland Heart Disease dataset (UCI Repository), contained 296 instances and 14 attributes after preprocessing. The dataset was analyzed using Weka software, and classification models were evaluated based on sensitivity, specificity, and accuracy. Notably, the SVM model achieved the highest accuracy of 84.12%, outperforming RIPPER (81.08%), ANN (80.06%), and Decision Tree (79.05%). The study highlights the effectiveness of SVM in heart disease prediction, demonstrating its superior performance among evaluated classifiers [64].

In 2013, Dhanashree S. Medhekar et al. proposed a machine learning model for the prediction of heart disease. The proposed framework utilized the Naïve Bayes algorithm for classification and prediction. The model aimed to categorize medical data into five risk levels: No, Low, Average, High, and Very High. The dataset, obtained from the Cleveland Heart Disease dataset (UCI Repository), contained 303 instances and 14 attributes after preprocessing. The dataset was splited into training and testing sets with varying instance distributions. Notably, the Naïve Bayes model achieved a maximum accuracy of 89.58%, depending on the number of training records. The study highlights the effectiveness of Naïve Bayes in medical data classification and heart disease prediction [65].