IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

Hyperparameter Optimization for Disease Detection and Analysis

Dr. Markad Ashok[1], Mr. Dawkhar Ashish[2], Mr. Chaudhari Kunal[3], Ms. Kanawade Tanvi[4], Ms. Kanawade Vaishnavi[5]
Department of Information Technology[1,2,3,4,5]
Amrutvahini College of Engineering, Sangamner, Maharashtra, India

Abstract: *The heart is crucial for living organisms, and detecting heart-related diseases necessitates accurate and precise monitoring. Cardiovascular disease is the primary cause of mortality across the world. Machine learning can assist in predicting heart disease survivors by converting large amounts of healthcare data into valuable insights for decision-making. This is a critical challenge in clinical data analytics. Various studies have identified important attributes that have a significant impact on predicting heart disease survivors. Machine learning can assist in uncovering these crucial attributes and assist healthcare professionals in anticipating a patient's survival and then adapting their care plan appropriately. As such, machine learning has great potential to improve patient outcomes and reduce healthcare costs associated with heart disease. Machine learning systems have shown potential in predicting and detecting cardiovascular disease (CVD) at an early stage, which can help mitigate mortality rates. Several research studies have utilized various machine learning techniques to identify CVD and determine the severity level of patients, yielding promising results. These approaches have the potential to assist healthcare professionals in improving patient outcomes and reducing the burden of CVD on society. This study proposes a method to address imbalance distribution in predicting patient status using the Synthetic Minority Oversampling Technique (SMOTE). Six machine learning (ML) classifiers were used and Hyperparameter Optimization (HPO) was employed to find the best hyperparameters. The results show that the proposed method improved the performance of the ML classifiers in detecting patient status. The findings suggest that the proposed approach could provide a valuable tool for improving diagnostic accuracy in medical applications. The model proposed in the study can assist doctors in identifying a patient's heart disease status, leading to early intervention and prevent mortality related to heart disease. By using this model, doctors can provide timely treatment and reduce the risk of heart disease-related complications. Implementing the model can help improve patient outcomes and reduce healthcare costs associated with heart disease management.*

Keywords: ML-Machine Learning, CVD-Cardiovascular Disease, SMOTE- Synthetic Minority Oversampling Technique, HPO-Hyperparameter Optimization

I. INTRODUCTION

Machine learning is a branch of AI in which software programs improve their accuracy in predicting outcomes without being explicitly programmed. This is possible through the use of algorithms that analyse historical data to predict future values. As the algorithm is trained on more and more data, its accuracy improves and it becomes better at accurately predicting outcomes. This has numerous applications in fields such as finance, healthcare, and marketing, and has the potential to revolutionize the way businesses operate. Machine learning is a crucial technology for enterprises, particularly in use cases like recommendation engines, fraud detection, spam filtering, malware threat detection, business process automation, and predictive maintenance. Machine learning helps businesses identify trends in consumer behaviour and streamline operations by automating processes. With the ability to handle large datasets, machine learning algorithms provide accurate predictions, enabling organizations to make informed decisions and achieve business goals more efficiently. In sum, machine learning is a powerful tool for improving efficiency, productivity and driving business success. Supervised learning is one of the four basic approaches in machine learning,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

along with unsupervised learning, semi-supervised learning, and reinforcement learning. In this approach, data scientists provide algorithms with labelled training data and specify the variables they want the algorithm to evaluate for correlation. The training data helps the algorithm to learn patterns and relationships between input variables and outputs. This approach is commonly used in applications that involve classification or regression problems. Supervised learning can assist in error reduction, and the accuracy increases as the amount and quality of training data increases. The increasing size of medical information systems in hospitals and medical institutions makes the process of extracting useful information increasingly difficult. Manual data analysis is no longer efficient, and computer-based methods are needed to perform efficient analysis. The integration of data mining in medical analysis can offer several advantages, such as heightened precision in diagnosis, lower expenses, and less manpower needed. This approach has been endorsed by evidence-based research.

The leading cause of mortality worldwide, according to the WHO, is cardiovascular disease. This encompasses a wide variety of conditions, including abnormalities in the heart's arteries, veins, and muscles. Due to the prevalence and severity of cardiovascular disease, accurate detection methods are vital. Machine learning algorithms have been utilized to forecast CVD using clinical datasets, but challenges arise from class imbalance and high dimensionality. Methods must be developed to address these limitations and improve the accuracy of CVD detection using machine learning techniques. Clinical datasets pose challenges due to class imbalance and high dimensionality, which reduce the efficiency and accuracy of machine learning methods. Feature selection has been studied to address these challenges. Studies have recently focused on developing decision support systems to tackle imbalanced datasets. One approach involves a balancing technique, and another has developed an HD prediction method utilizing DBSCAN and hybrid synthetic minority over-sampling technique-edited nearest neighbours. These techniques have shown promise in improving the accuracy of predictions on imbalanced datasets, which is essential for various applications. Such advancements could have significant impacts on decision-making processes, particularly in the medical and financial sectors, where data is often highly imbalanced. SMOTE technique was used by the author to balance the dataset without feature selection. A study used SMOTE to balance data distribution while also utilizing extremely randomized trees (ET) on selected parameters for predicting patient survival by leveraging the importance ranking of random forests.

An algorithm that efficiently predicts heart attacks has been proposed, alleviating the need for costly feature engineering. However, the imbalanced nature of the dataset may still pose challenges to accurate classification. The proposed algorithm utilizes end-to-end learning to minimize the cost associated with feature engineering in classification tasks. It does not require any pre-processing of data, making it efficient and cost-effective. The algorithm also addresses the issue of imbalanced datasets through an effective approach. Overall, this algorithm provides a simple and effective solution for classification tasks while reducing the burden of feature engineering and addressing dataset imbalances.

A number of factors which affect hearth health like blood pressure, level of cholesterol, creatine, etc., so it makes it difficult to diagnose. Analysed different factors that cause heart disease and identified controllable factors like alcohol usage, smoking, diabetics, high level of cholesterol, and limited physical activity. Electronic health records (EHRs) are valuable tools for both clinical and research purposes. Machine learning-based expert systems have proven to be effective in diagnosing cardiovascular disease (CVD), where minor errors in physical examinations could have significant consequences. EHRs help to ensure accurate and timely diagnoses, potentially saving lives. The use of machine learning in medical diagnosis is an exciting area of research that holds great promise for improving patient outcomes. Data mining plays an immense role in extracting useful information from big data. It is widely used in almost every field of life like medicine, engineering, business, and education. Data mining is used to explore the data to extract the hidden crucial decision making information from the collection of the past repository for future. A variety of ML algorithms have been used to understand the complexity and non-linear interaction between different factors by decreasing the error in prediction and factual outcomes. Due to ever increasing medical data, we need to leverage on machine learning algorithms to assist medical healthcare professionals in analysing data and making accurate and precise diagnostic decisions. Medical data mining makes use of various classification algorithms to predict the incidence of cardiovascular disease (CVD) and the risk of death due to heart attack in patients. The algorithms analyse data from various sources such as medical records, genetics, lifestyle factors, and imaging tests to generate accurate predictions. The predictions help physicians make informed decisions on diagnosis, treatment, and prevention of CVDs,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

thereby improving patient outcomes. The classification algorithms commonly used in medical data mining include decision trees, logistic regression, support vector machines, K-nearest neighbors, random forests, and neural networks.

II. LITRETAURE SURVEY

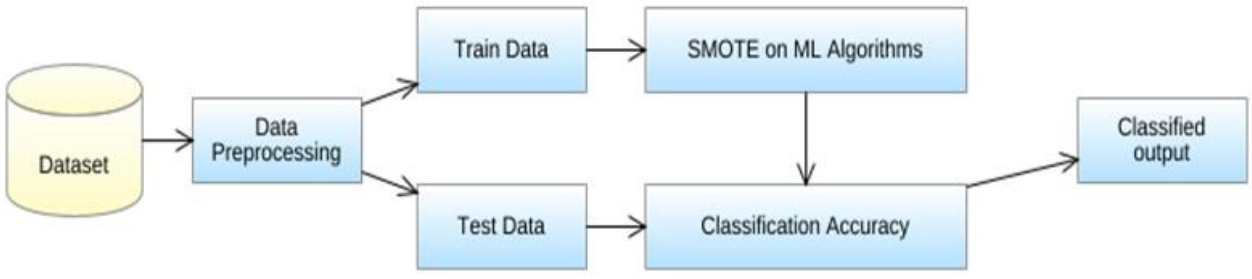
Abdellatif, H. Abdellatef, J. Kanesan, C. -O. Chow, J. H. Chuah and H. M. Gheni, An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods: This study utilizes frequent pattern growth association mining to extract information directly from electronic patient records and generate strong association rules. This eliminates the need for manual data extraction and can reduce the time and resources required for analysis This paper utilizes classification techniques and the data mining tool WEKA to predict heart disease in male patients. It provides an in-depth look at coronary heart disease, including common types and risk factors. The authors' use of these techniques has the potential to aid in the early identification and treatment of heart disease, ultimately improving patient outcomes.

K. S. K. Reddy and K. V. Kanimozhi, Novel Intelligent Model for Heart Disease Prediction using Dynamic KNN (DKNN) with improveda accuracy over SVM: The aim of this study is to develop a novel intelligent model for predicting heart disease using Dynamic KNN and compare it to Support Vector Machine (SVM). Dynamic KNN is a straightforward algorithm used to predict heart disease in 92 patients. The study's materials and procedures employ two machine learning techniques. The research's main objective is to forecast heart disease using a dynamic KNN algorithm and compare the results to SVM to determine which approach is more effective. Dynamic KNN was found to have an accuracy of 84.44% for predicting heart disease, while Support Vector Machine only had an accuracy of 67.21%. These results suggest that Dynamic KNN is significantly better at predicting heart disease than SVM. Therefore, Dynamic KNN may be a more reliable method for predicting novel heart disease.

G. S. Reddy Thummala and R. Baskar, Prediction of Heart Disease using Machine Learning Algorithms Application of Support Vector Machine Based on Particle Swarm Optimization in Classification and Prediction of Heart Disease: In this study, real-world electronic health record data related to congestive heart disease was used to predict the occurrence of the disease. The researchers utilized one-hot encryption and word vectors to model the diagnostic events and effectively predict coronary failure. This approach has the potential to improve healthcare outcomes by allowing for early intervention and treatment of heart disease.

Prediction of Heart Disease using Decision Tree in Comparison with KNN to Improve Accuracy: The study aims to enhance the forecasting accuracy of cardiac disease by utilizing machine learning algorithms and comparing their performance to that of the K-Nearest Neighbor method. The research involves exploring two groups- Decision Tree and K-Nearest Neighbor. The aim is to implement the irregular forest method to improve cardiac disease prediction. The findings could offer valuable insights into the application of advanced machine learning algorithms for enhancing medical forecasting accuracy. After conducting 20 iterations on each strategy using a dataset of 1700 records, the decision tree proved to be the most effective method with a repeated measures power of 80%. This method was able to accurately classify records with the highest level of accuracy. The other strategies also showed promise and may be useful in specific contexts, but did not perform as consistently as the decision tree method. This experiment demonstrates the importance of testing and comparing multiple strategies when working with large datasets to ensure optimal model performance.

III. METHODOLOGY



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

Step 1: The data is collected from the database, cleaned and pre-processed, including removing missing data, label grouping, and data normalization using min-max.
Step 2: After data pre-processing, the data is split into training and testing.
Step 3: The training data is then entered into the SMOTE technique for data distribution balancing
Step 4: The ML model is built and evaluated using recall and AUC; when the ML model reaches the optimal performance.
Step 5: After the best ML hyperparameters is selected, the ML model enters the criteria for updating SMOTE hyperparameters. If the ML model satisfies the stopping criteria, it will proceed to the next step.
Step 6: In the last step, the model is evaluated with existing models.

IV. PROPOSED WORK

Decision Tree:Decision Tree is a powerful algorithm used for solving classification and regression problems. It is a tree-structured classifier where the internal nodes represent the features of a dataset, and the branches represent the decisions. It can handle both categorical and continuous variables, making it efficient for real-world problems. Decision Tree is popular because of its visualization capability as it can be easily understood, and the logic behind the decisions made by the model can be easily explained. This algorithm is useful in various applications such as finance, healthcare, marketing, and many more. A Decision tree consists of two nodes, Decision and Leaf nodes. Decision nodes contain multiple branches and are used to make decisions through testing. Leaf nodes do not contain further branches but represent the output of those decisions. The tree structure is formed by connecting these nodes using edges. The decision tree is a commonly used method for decision-making and classification in various fields, including machine learning, finance, and medicine. It is easy to interpret and visualize, making it a popular choice for data analysis.
Linear Regression: Using ML algorithm linear regression and SVM are applied on a group different of clusters. Linear regression separate the data point from a single line. We get accuracy improved of the data. It is used to estimate real values(cost of houses, number of calls, total sales etc) based on continuous new variable(s). Here, we build up connection among free and ward factors by the fitting a best line. This best fit line is known as relapse line and spoken to by a direct condition and represented by a linear equation;

$$Y = a * X + b$$

These coefficients a and b are determined dependent on limiting the aggregate of squared contrast of separation between information focuses and relapse line.
Random Forest: The random forest algorithm is a powerful tool in machine learning that uses ensemble learning to solve complex regression and classification problems. It consists of many decision trees, each of which makes a prediction based on a different subset of input variables. This approach helps to reduce overfitting and increases the accuracy of the results. Random forests have proven to be effective in many applications, including image and speech recognition, fraud detection, and medical diagnosis. They are widely used in industry and academia, and continue to be an active area of research and development.
XGBoost: XGBoost is a powerful machine learning algorithm known for its ability to handle large datasets and deliver exceptional performance in classification and regression tasks. With its extreme gradient boosting technique, XGBoost has become one of the most widely used algorithms today. It is a popular choice for companies and data scientists alike to make predictions and solve complex data-driven problems. XGBoost is a powerful library that is used for training machine learning models in a scalable and efficient manner. It uses a technique called ensemble learning to combine the predictions of multiple weak models to produce a single, stronger prediction. This makes it an excellent choice for applications where accuracy is crucial, such as in the financial industry or in healthcare. Additionally, XGBoost is designed to work with distributed computing systems, allowing it to handle large datasets with ease.
SMOTE is an effective oversampling technique for addressing class imbalance by generating synthetic samples for minority classes. The method randomly replicates minority class instances to achieve a more balanced dataset that can improve classification accuracy. SMOTE has proven to be an efficient solution to problems of data imbalance, making it a widely used tool in machine learning. SMOTE is an oversampling technique that generates new minority instances by creating synthetic samples between existing minority instances. This approach is commonly used in medicine to address imbalanced class datasets. SMOTE helps to prevent bias in data analysis and improves the accuracy of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

classification models by providing a balanced dataset. By synthesizing new instances, SMOTE allows the algorithm to learn from more diverse examples of the minority class that might be insufficiently represented in the original dataset, leading to better predictions. SMOTE is a technique used for generating synthetic data of the minority class by utilizing nearest neighbors and Euclidean distance. By creating new samples based on the original characteristics, SMOTE is able to increase data instances that closely resemble the original data. This approach is widely considered to be effective and reliable for addressing class imbalance in machine learning models.

SMOTE Algorithm

SMOTE is an algorithm for data augmentation that creates synthetic data points based on the original data. Unlike simple oversampling, SMOTE generates new data instead of just duplicating existing data. The synthetic data points are created by finding the nearest neighbors of each point in the original dataset and interpolating between them. This approach can help to address class imbalance problems in machine learning, where one class of data is underrepresented compared to another. By creating additional synthetic data points for the underrepresented class, SMOTE can help improve the accuracy of predictive models.

The **SMOTE algorithm** works as follows:

You draw a random sample from the minority class.

For the observations in this sample, k closest observations to a given data point will be identified.

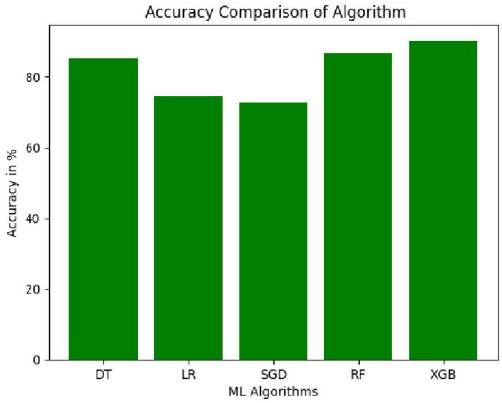
After finding the nearest neighbors of a data point, we can identify the vector between the current data point and the selected neighbor.

You multiply the vector by a random number between 0 and 1.

To generate new data points, synthetic data can be added to the existing data points.

V. EXPERIMENTAL RESULTS

Classifiers trained were compared to see which performed better on datasets. On evaluation metrics, some classifiers performed well, while others performed poorly. The tree-based ensemble models most improved using HB since those models contain more hyperparameters to be optimized with a more extensive search space area, which significantly affects the classifier's performance. In contrast, the regression-based models didn't improve much since those models depend more on data distribution.



XGBoost classifier achieved the highest accuracy by 90.19%.Based on this result assessment and evaluation, the present study on the CVD detection prediction model delivers superior outcomes than the prior work in terms of accuracy, f1-score, and MCC.

VI. CONCLUSION

The proper filters were utilized to improve the validity and rationality of the dataset. The combination of SMOTE and ML models is expected to improve prediction accuracy in the proposed model, however, pre-processing may take more time as a limitation of this approach. The experiment revealed that the use of tree-based models delivered better outcomes with higher quality results. Additionally, the Hyperparameter optimization technique had a significant impact on enhancing the accuracy of the models. Thus, SMOTE by hyperparameter achieved the highest accuracy for binary

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 1, May 2023

and multi-class problems on dataset. The accuracy are compared with other existing methods. This study aims to enhance the existing methodologies by introducing a novel and distinct approach for model creation. Its primary objective is to formulate a model that is practical and applicable in real-life scenarios. The study's central theme is to advance the existing understanding and to establish a more simplified, user-friendly, and relevant model. The framework will be focused on delivering an efficient and effective solution that can cater to the needs of the industry and academia. Through this study, we envision providing a comprehensive and practical solution that can benefit various stakeholders.

REFERENCES

[1]. A. Abdellatif, H. Abdellatef, J. Kanesan, C. -O. Chow, J. H. Chuah and H. M. Gheni, ”An Effective Heart Disease Detection and Severity Level Classification Model Using Machine Learning and Hyperparameter Optimization Methods,” in IEEE Access, vol. 10, pp. 79974-79985, 2022, doi: 10.1109/ACCESS.2022.3191669

[2]. K. S. K. Reddy and K. V. Kanimozhi, ”Novel Intelligent Model for Heart Disease Prediction using Dynamic KNN (DKNN) with improved accuracy over SVM,” 2022 International Conference on Business Analytics for Technology and Security (ICBATS), 2022, pp. 1-5, doi: 10.1109/ICBATS54253.2022.9758996.

[3]. T. Xue and Z. Jieru, ”Application of Support Vector Machine Based on Particle Swarm Optimization in Classification and Prediction of Heart Disease,” 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), 2022, pp. 857-860, doi: 10.1109/ICSP54964.2022.9778616.

[4]. G. S. Reddy Thummala and R. Baskar, ”Prediction of Heart Disease using Decision Tree in Comparison with KNN to Improve Accuracy,” 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2022, pp. 1-5, doi: 10.1109/ICSES55317.2022.9914044.

[5]. R. T. Selvi and I. Muthulakshmi, ‘‘An optimal artificial neural network based big data application for heart disease diagnosis and classification model,’’ J. Ambient Intell. Humanized Comput., vol. 12, no. 6, pp. 6129–6139

[6]. G. Bazoukis, S. Stavrakis, J. Zhou, S. C. Bollepalli, G. Tse, Q. Zhang, J. P. Singh, and A. A. Armoundas, ‘‘Machine learning versus conventional clinical methods in guiding management of heart failure patients—A systematic review,’’ Heart Failure Rev., vol. 26, no. 1, pp. 23–34.

[7]. A. Makhlouf, I. Boudouane, N. Saadia, and A. R. Cherif, ‘‘Ambient assistance service for fall and heart problem detection,’’ J. Ambient Intell. Humanized Comput., vol. 10, no. 4, pp. 1527–1546.

[8]. M. Chen, S. Gonzalez, V. Leung, Q. Zhang, and M. Li, ‘‘A 2G-RFIDbased e-healthcare system,’’ IEEE Wireless Commun. Mag., vol. 17, no. 1, pp. 37–43.