[1]Monali Gulhane

[2]T. Sajana

[3]Nitin Rakesh

Optimized Approach for Heart Disease Prediction by using Enhancing Diagnostic Method through Meta-Features

Abstract: - In today's world, pollution is worsening, and bad habits like not eating regularly, eating lots of junk food, and not exercising enough are becoming more common. This can cause health problems, and the global need for the detection of heart diseases is increasing. The American Heart Association, in 2023, has given an update that heart disease is the leading cause of death. Implementing machine learning models has given significant results, but due to the limitations of requirement balanced data, repeat training model complexity has led to unreliable results in some cases. Hence, the proposed model overcomes the limitations by comparing machine learning models for cardiovascular diseases. Performance of proposed work is evaluated with the base models KNN and DT with and without smooth. Thus, the comparison included increased accuracy in DT, but the proposed model GBM as a meta learner has led the performance metric with 92% accuracy, with recall of 0.89% and F1 score of 0.86%. Thus, the proposed approach has achieved the highest accuracy in cardiovascular diseases using meta-subjects. Future developments for the research will focus on applying the model to a larger dataset and analysing cases according to complex machine learning models.

Keywords: Machine learning, disease predictions, k nearest neighbour, decision tree, meta learner, meta features.

I. INTRODUCTION

Globally, cardiovascular diseases are the foremost cause of mortality, per the WHO. Identification of cardiovascular disease (CVD) can be challenging due to the presence of numerous contributing factors, including but not limited to hypertension, hyperlipidaemia, diabetes, and irregular heart rate. Occasionally, symptoms of CVD may differ between males and females. For instance, chest pain is more prevalent in male patients, whereas female patients may also experience vertigo, excessive fatigue, and shortness of breath in addition to chest distress. Researchers have been investigating an extensive array of methodologies to forecast cardiovascular ailments. However, early disease prognosis remains inefficient for a variety of reasons, which include but are not limited to execution time, intricacy, and approach accuracy. Because of this, accurate diagnosis and treatment have the potential to preserve numerous lives. Many factors, including blood pressure, cholesterol levels, creatine, and others, influence cardiac health, complicating the diagnostic process. The authors in the research for heart disease prediction have identified controllable risk factors for heart disease, including tobacco use, alcohol consumption, diabetes, elevated cholesterol levels, and insufficient physical activity. Electronic health records, or EHRs, are becoming valuable tools for clinical research. There is a possibility that the physical examination may contain some inaccuracies, which, if they lead to cardiac disease, could ultimately be fatal. The utilization of expert machine learning-based systems in diagnosing heart diseases leads to a reduction in the mortality ratio. The regular methods for predicting heart diseases include traditional machine learning algorithms using CSV data. However, the problem of model implementing the model on the imbalanced data persisted again and again. A highly efficient algorithm proposed in our research has been developed to accurately forecast the occurrence of heart attacks using a specific dataset. A significant drawback of the current research was the requirement for extensive and expensive feature engineering in the classification process. In addition, the unbalanced nature of the data set can negatively impact the overall effectiveness of the classification algorithm, affecting its accuracy and reliability. The proposed algorithm focuses on reducing the expenses related to feature engineering. This approach relies on end-to-end learning, where pre-processed data is directly used for classification without feature engineering. In addition, we delve into the imbalanced nature of the provided dataset and contribute to a practical method to enhance the dependability of the classification outcomes.

[1] 1Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

2Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

3Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India

Corresponding author: monali.gulhane4@gmail.com

Thus, the research to assist healthcare practitioners by creating machine learning algorithms for predicting the survival of heart disease patients, which give accurate results, also solve the problem of the imbalance data. We have also tried to repeatedly reduce the time spent training the data for the new algorithms. Instead, we have implemented the meta feature, which is used as input to the meta learner model. The research includes implementing the KNN with smote, KNN without smote, DT with smote, and DT without smote, which gives the final meta learners to the system. The data are divided into three sets: train, validate and test. Further, we have two output outputs: I am from test data along with machine learning algorithm: KNN with smote, KNN without smote, DT with smote, DT without smote, and output II is output from again test data with the meta learner model. The Synthetic Minority Oversampling Technique (SMOTE) solves the class imbalance issue. The key contribution of the research is as follows:

- End-to-End Learning Approach: The research proposes a novel approach that passes the traditional implementation of machine learning to learn and then test the data; it uses the combination of machine learning to analyse the data for accurate prediction by directly using the process data, this approach is known as end-to-end learning.
- Meta Feature implementation with meta learner models: Using the meta-learning approach in combination with the machine-learning approach reduces the computational cost and time required to retrain the model with the proposed algorithm.
- The proposed approach also uses the Synthetic Minority Oversampling Technique (SMOTE) with a combination of machine learning algorithms (KNN and DT). This overcomes the problem of using only one algorithm that can give reliable results on unbalanced data.

The paper explains the machine learning algorithms for heart diseases predictions using meta learner’s model, the article in section II describes the in-depth literature review, for analysis of traditional machine learning for diseases prediction along with its limitations. Section III of the paper explains the proposed model with the implementation of the meta features and meta learners. Section IV describes the result analysis with the case comparison of all the models. Section V explains the conclusion with the results achieved for the implemented model.

II. LITERATURE REVIEW

The literature review explains the in-depth survey about machine learning and implementing the meta-learner's model. Numerous research studies have examined using SMOTE in conjunction with conventional machine learning approaches to predict heart disease. After using SMOTE, Bouqentar et al. assessed the accuracy of the logistic regression (LR) method and naive Bayes (NB) classifier on the Cleveland and Hungarian datasets, obtaining 92% and 90% of the results, respectively [1]. Yang and Guan presented a framework for predicting cardiac disease using the smote-xgboost algorithm. Research explains eight classification models, including naive Bayes (NB). When forecasting the prognosis of cardiac patients, its Decision Table/Naive Bayes hybrids classifier (DTNB) had the best accuracy, coming in at 87.08% [2]. The optimal predictors of bad outcomes were identified, patients at risk of adverse effects were precisely stratified, and the ability to differentiate the efficacy of adverse results in heart failure patients was successfully enhanced by the combination of SMOTE+ENN and cutting-edge machine learning techniques [3].

To address the imbalanced data, the suggested study employs a synthetic minority oversampling technique (SMOTE) [4]. The suggested method eliminates the need for feature engineering to classify the provided dataset.

The results of the UCI machine learning cardiovascular database are evaluated and reviewed using machine learning methods. The proposed method achieved the best accuracy, and the classification using random forest scored 96.72% and the gradient scored 95.08%[5]. Conventional methods need to be modified for analysis and prediction. Deep learning methods require large datasets, which are not available in clinical or scientific research [6]. The expected results show improved class accuracy. When analysing the data for sensitive records, the mock data set created plays an important role in improving class mathematics. This study[7] demonstrated the feasibility of using machine learning techniques to accurately predict cardiovascular disease. Appropriate model selection including random forest, logistic regression, decision tree, and KNN algorithm provided robust and reliable prediction The aim of the study was to develop ML models for prediction of cardiovascular disease a relevant factors will be used The study explains that this study's UCI cardiac prediction benchmark dataset includes 14 cardiac-specific datasets. Research also sought to reveal relationships between database attributes using traditional machine learning.

Cardiovascular disease prognosis is an important endeavour that requires the development and exploration of a prognostic method to prevent cardiovascular disease and inform patients before the condition occurs [9]. The aim of this study is to determine a method that provides excellent accuracy in cardiovascular disease [10]. The aim of this study is the early diagnosis of cardiovascular disease. To predict cardiovascular disease, we trained the model using various methods on the training data set, including logistic regression, k nearest neighbors (kNN), decision trees, and random forest model and then, testing the data set checked its accuracy. The random forest method fits the data well, with an accuracy of 88.16%. The study explains that the accuracy of traditional machine learning techniques can be very effective in detecting cardiovascular diseases. Thus, meta-studies are an alternative approach defined by the research study; meta-studies have demonstrated the ability to predict cardiovascular risk and increase classification accuracy [11]. Machine learning techniques such as ensemble learning and meta-classifiers have been used to increase the accuracy of cardiovascular disease prediction [12].

Furthermore, a meta-learning framework was created to learn stacked Restricted Boltzmann Machine (RBM) models for the classification of heart disease, resulting in state-of-the-art accuracy [13]. These results indicate that meta-learning techniques might help predict and categorize cardiac disorders. From the survey, we have analyzed the limitations as shown in the Table below; the research has achieved results to overcome the limitation requirement of hyperparameter tunning, the requirement of large datasets, and the complexity of combining the decision tables.

Table1: Survey Analysis

Method Used	Training Limitations
SMOTE with LR and NB	Understanding balance in synthetic sample generation; expertise in synthesized data assessment.
SMOTE-XGBoost and DTNB	Computational intensity for XGBoost tuning; complexity in combining decision tables and NB.
SMOTE+ENN with Advanced ML Techniques	Expertise in oversampling and noise removal; high computational resource needs.
SMOTE for Imbalanced Data	Knowledge in effective SMOTE application; evaluating synthetic data's impact.
Random Forests and Gradient Boost	High computational resources for training; essential hyperparameter tuning knowledge.
Deep Learning Techniques	Large datasets and significant computational resources needed; deep learning expertise.

III. PROPOSED MODEL

The research explains the four classes of heart disease prediction. The proposed model for the meta-learning in system synthesis is the process used to train a secondary model (also known as the meta-learner) to merge base model predictions, as shown in Figure1. The base models in the proposed model implemented are KNN and DT, with and without SMOTE added. The dataset is pre-processed to check for missing values and checking for balanced and imbalanced data. The base model is implemented on the validated dataset, and the outcome of this process is "Meta Features" These meta-features are given as input to the "Meta Learner" model. The implemented meta-learner model in this research is GBM (Gradient Boosting Machine). The primary objective of this meta-learner, which may be either a Gradient Boosting Machine or a lightweight neural network, is to investigate the optimal approach for integrating the predictions generated by the base models. Using this meta-learning degree, the version attains a refined and advanced predictive capability, transcending the capabilities of the individual base model. Output I of the base models for KNN without smote and with smote, DT with smote and without smote is compared using hard voting with Output II of the meta learner model. This approach effectively solves the class imbalance challenges and utilizes several prediction models' advantages.

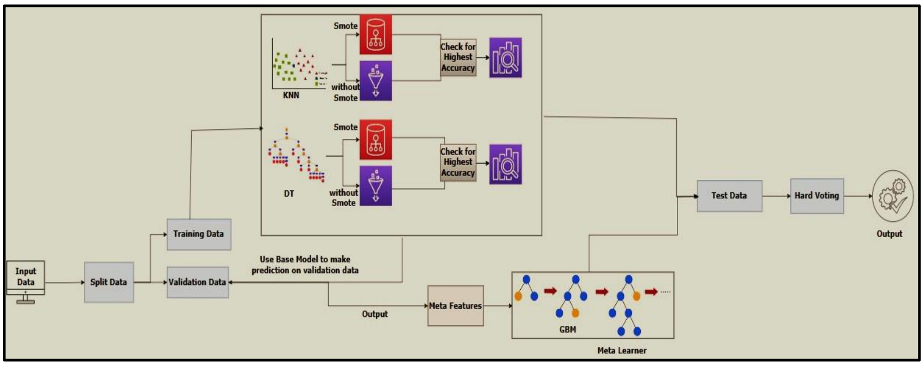

Figure 1: Flow Chart of Proposed Model

It incorporates a meta-learning layer that enhances the overall performance of these models by optimizing their collective insights. The presented approach represents a significant advancement in predictive modelling for cardiac disorders, offering an excellent opportunity for improved diagnostic precision and enhanced patient treatment. The proposed model is explained in Figure 1 and Algorithm I.

3.1 Dataset:

The dataset used is the BMC Medical Information Technology and Decision-making, 20, 16 (2020) Dataset from Davide Chicco and Giuseppe Jurman; it is applied using machine learning, used to predict survival for individuals with heart failure or strokes.

- Size: It contains 920 rows (entries) and 14 columns (variables), occupying around 100.8 kilobytes of memory.
- Columns with the data can be described in Table 2 and Table 3

Non-Null Values: There are 920 non-null values across most columns, and all the features explained are essential for predicting heart diseases.

```
# Algorithm 1: Heart Disease Prediction with Meta-Learning
Input:- Pre-processed heart ailment dataset.
Output: Performance metrics of the meta-learner model are at the check set.
1. Load and preprocess the heart sickness dataset.
2. Split the dataset into training, validation, and check sets. For each aggregate (KNN with/without SMOTE, DT with/without SMOTE):
    a. If SMOTE is implemented, oversample the minority elegance in the training set using SMOTE.
    b. Train the model (KNN or DT) at the (possibly SMOTE-oversampled) training set.
3. Generate Meta-Features at the Validation Set:
    a. Use each of the educated base models to predict outcomes on the validation set.
    b. Collect these predictions to shape a new dataset of meta-functions.
4. Train Meta-Learner:
    a. Use the meta-feature dataset (from step 4) as input and the real validation set results as output to train the meta-learner version.
5. Generate Meta-Features at the Test Set:
    a. Use the trained base models to predict results at the check set.
    b. Collect those predictions to form a meta-feature dataset for the check set.
6. Final Prediction via Meta-Learner:
    a. Use the meta-learner model to expect effects at the check set primarily based on the test set meta-functions generated in step 6.
7. Evaluate Performance:
    a. Compare the meta-learner's very last predictions against the actual effects of the take a look at the set.
    b. Calculate and document performance metrics (e.g., accuracy, F1 rating, ROC AUC).
End Algorithm
```

Table 2: Details of Dataset

age	sex	cp	treetops	chol	FBS
25	Male	Typical angina	130	200	> 120 mg/dL
30	Female	Atypical angina	110	250	<= 120 mg/dL
35	Male	Nonanginal pain	140	300	> 120 mg/dL
40	Female	Atypical angina	120	270	<= 120 mg/dL
45	Male	Typical angina	150	220	> 120 mg/dL

Table 3: Details of Dataset Conti…

age	restecg	thalach	exang	old peak	slope	ca	thal	num
25	Normal	170	Yes	2	Upsloping	0	Normal	1
30	ST-T abnormality	150	No	1	Flat	1	Normal	2
35	Normal	190	No	3	Downsloping	2	Fixed defect	3
40	Normal	160	Yes	2	Flat	1	Normal	4
45	Abnormal	180	Yes	4	Upsloping	0	Reversible defect	5

1.2 Exploratory Data Analysis:

3.2.1 Analysis I:

The statistics in the Figure 2 reveal that there are more males than women in the sample. Atypical angina, or a specific sort of chest pain, seems to be the most prevalent complaint among patients. High blood sugar levels in this patient group may be associated with coronary artery disease since fasting blood sugar levels were measured. Abnormal ECG readings suggest that some individuals may already have cardiac damage. The exercise stress test indicated that several individuals suffered angina during physical activity, which might indicate a problem with the coronary artery function. Differences in ST segment slope with exercise may reflect different severity levels of heart disease throughout the patient population.

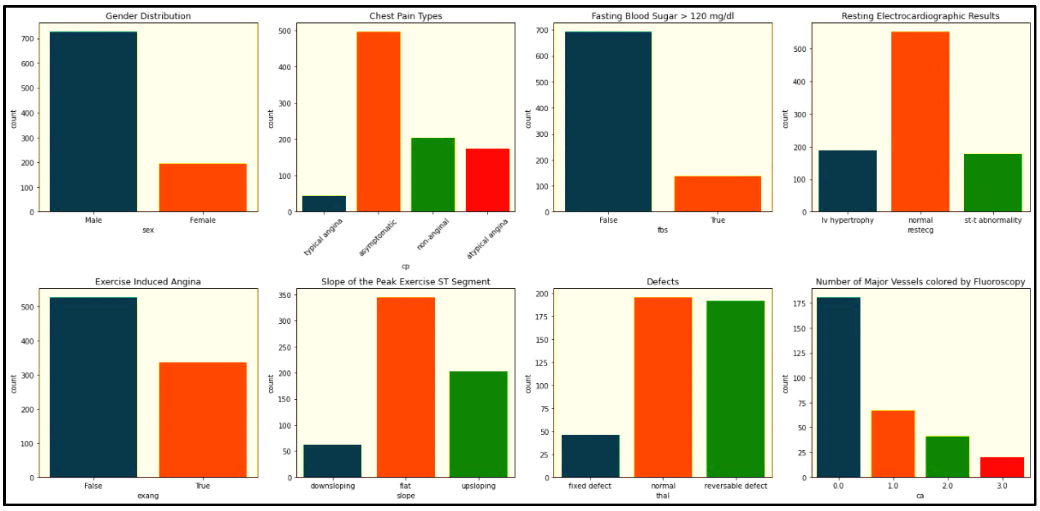

Figure 2: Database Analysis I

3.2.2 Analysis II:

The age-cholesterol plot in the Figure 3 suggests that cholesterol levels tend to rise with age, particularly up to around 65 years old. The age-blood pressure plot indicates a positive relationship, where blood pressure increases as people age. However, the age-depression plot does not show a clear trend, making it less definitive about the connection between age and depression.

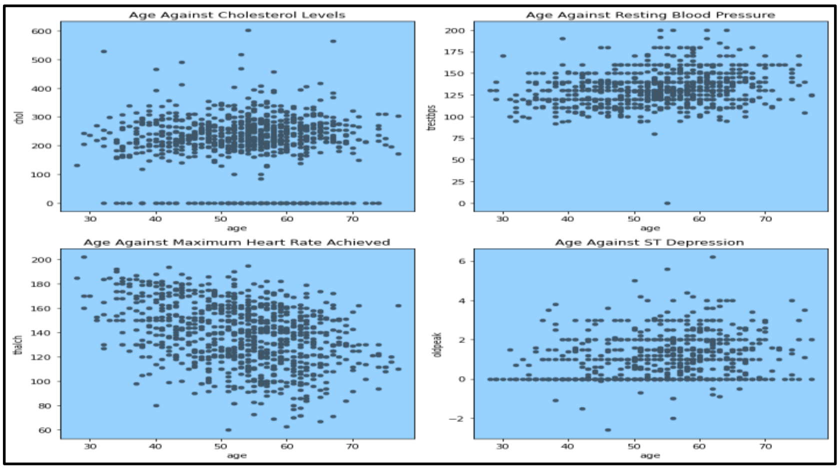

Figure 3: Database Analysis II

3.2.3 Analysis III:

We have analyze in the Figure 4 that if cholesterol levels rise from low to medium to high, it may be feasible to identify whether there is an increase or decrease in heart rate and blood pressure in each group of individuals (males and females). The statistics would offer particular values for beats every minute and blood pressure based on cholesterol level.

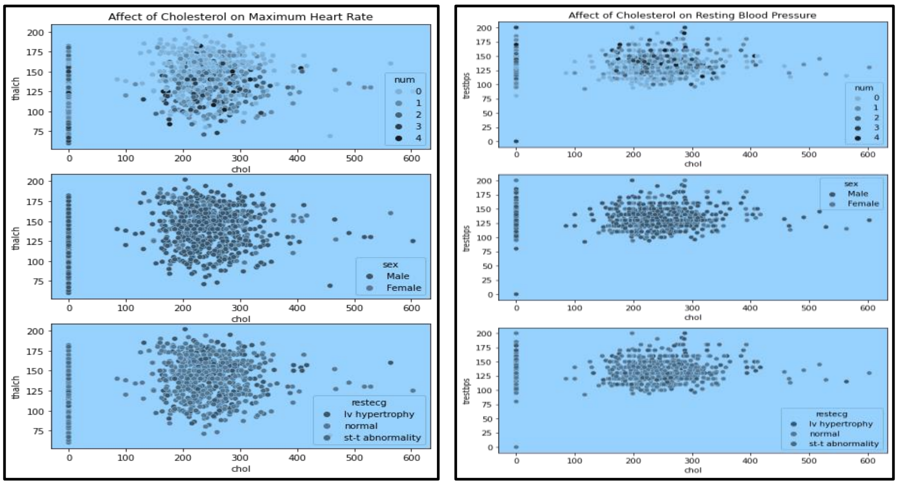

Figure 4: Database Analysis III

3.4 Preprocessing:

3.4.1. ***Checking for Missing Value:***

The data analysis gives insight into the missing values in the dataset, which is explained in Table 4 The missing values are handled in the following steps

Step 1: Drop columns having a large number of missing values.

```
heart_df.drop(labels=['ca', 'thal', 'slope'], axis=1, inplace=True)
```

Step 2: Restructure the data types

```
heart_df = heart_df.astype({'sex':'category', 'cp':'category', 'fbs':'bool',
'restecg':'category', 'exang':'bool'})
```

Table 4: Missing Values

Name of Feature	Missing Values	Percentage
treetops	59	6.41%
chol	30	3.26%
FBS	90	9.78%
restecg	2	0.22%
thalach	55	5.98%
exang	55	5.98%
old peak	62	6.74%
slope	309	33.59%
ca	611	66.41%
thal	486	52.83%

3.4.2. Checking for balance data:

The dataset used in the research is not evenly distributed; the distribution of class on the dataset can be explained in Figure 5,

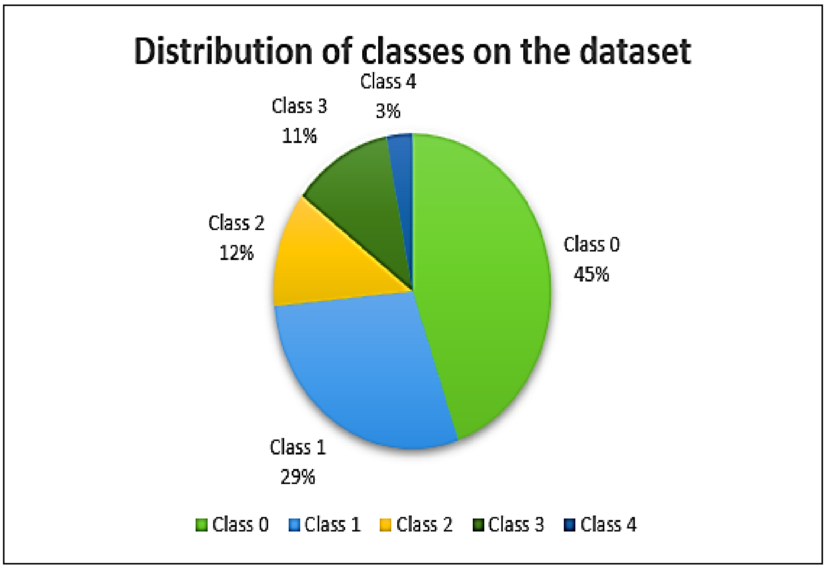

Figure 5: Analysis of Imbalance Dataset of Davide Chicco and Giuseppe Jurman

- **Class 0:** Indicate the absence of heart disease in the distribution of the dataset. It covers 45% of the data.
- **Class 1:** Indicates heart disease with slight severity; it covers 29% of the data.
- **Class 2:** Indicate the moderate form of heart disease; this class covers 12% of the data.
- **Class 3:** Indicates advanced phase of coronary artery diseases, which is more severe than moderate; this dataset covers 11% of the data.
- **Class 4:** Refers to the high severity of the cases; this dataset covers only 35 of the data.

As per the observation, the dataset is imbalanced; if the model is applied, the chance of false prediction increases; hence, it is necessary to balance the dataset. We have applied the SMOTE technique to balance the dataset with each implemented model.

3.5 Model Implemented

The analysis of multiple classes for predicting heart disease in this research significantly impacts studying heart diseases at early stages. The model implemented is as follows:

3.5.1 KNN:

The K-Nearest Neighbours (KNN) method is a standard machine-learning approach for regression and classification issues. It is based on the assumption that related data points usually have equivalent labels or values. The mathematical modified representation of the KNN is given in equation 1. This modified representation is an alternative to assigning the same weight to all k closest neighbours, providing weights inversely proportionate to distance. Closer neighbours have a more substantial effect on the forecast.

$$Prediction_R = \frac{\sum_{i=1}^{k} w_i * y_i}{\sum_{i=1}^{k} w_i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where, $w_i = \frac{1}{d(x1+x2)^2 + \in}$ to avoid division by zero

PredictionR: Output for KNN

wi : The weight associated with the ith nearest neighbour.

yi : The output value (e.g., class label or continuous value) of the ith nearest neighbour.

ith nearest neighbour.

k: The total number of nearest neighbors considered.

d(x1+x2): This represents the distance between the data point being predicted

d is a function that calculates the distance between two points (e.g., Euclidean distance).

ϵ: A small constant added to the denominator to avoid division by zero.

3.5.2 DT:

Decision Trees (DT) are supervised learning algorithms that may be used for classification and regression. They are named "decision trees" as the model employs a tree-like structure or model of choices and their potential outcomes. For implying the DT for the given dataset, the formula implemented is given in equation 2 for heart prediction.

$$HeartDiseasePrediction(x, Tree)\begin{cases} leafnodeprediction & if\ leaf\ node\ reached \\ Proceed\ based\ on & patient\ data\ and\ node\ decision\ criteria \end{cases} \ldots\ldots\ldots\ldots\ldots(2)$$

Where x is the patient data

3.5.3 Meta Features

For the dataset used in this research, we have dealt with four classes to check the severity of the heart diseases; for this analysis, we have two major base classes with two sub-base classes such as KNN with smote, KNN without smote, DT with smote and DT without smote, we have used the class labels directly as the meta-features.

Implementation of Class Labels Directly as Meta-Features:

Each base model is predicted as a class label for all instances in the set of validation datasets, where the class label is one of the four probable classes. In this state, we have four predictions (one from all four models), resulting in 4 meta-features.

a. Size of Matrix for Meta Feature: for N * 4, here, N indicate the number of instances in the validation set.

b. For each row: It represents the class predicted from labels from the base four models for the one instance

3.5.4 GBM as Meta Learner Model

The GBM, as meta learners have stated, corrects the base model's predictions by effectively combining them into a final prediction that targets to be more accurate than other individual base model predictions. The mathematical representation for this model is as shown in equation 3.

$$y = GBM([Model1(x), Model2(x) \dots\dots. Model3(x)]) \dots\dots\dots\dots\dots (3)$$

Where y is the prediction from the meta-learner

Model$i(x)$ indicate the prediction of the i-th base model for the input features n is the total number of base models used

3.5.5. Hard Voting

We have two outputs in the research for implementing meta-features with meta-learners and base models (KNN and DT). Hence, the research includes the complex voting ensemble to check the models directly and accurately.

IV. RESULT ANALYSIS:

4.1 Experiential Analysis

This research divides the dataset into three sets based on the train, test, and validate ratio. The experiment is done by changing the ratio to analyze the per analyzing the proposed model; this experimental observation is shown in Table 5.

Table 5: Experiential Analysis on Davide Chicco and Giuseppe Jurman Dataset

Case	Split Ratio	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Case 1	70% / 15% / 15%	KNN with SMOTE	0.87	0.86	0.87	0.85
		DT with SMOTE	0.88	0.89	0.88	0.88
		GBM Meta-Learner	0.92	0.92	0.89	0.86
Case 2	80% / 10% / 10%	KNN with SMOTE	0.86	0.84	0.83	0.81
		DT with SMOTE	0.88	0.87	0.85	0.89
		GBM Meta-Learner	0.89	0.88	0.88	0.88
Case 3	60% / 20% / 20%	KNN with SMOTE	0.81	0.81	0.81	0.81
		DT with SMOTE	0.77	0.75	0.79	0.77
		GBM Meta-Learner	82	81	83	82

The above table indicates that case 1 has improved results on the meta-learner GBM model compared to both base models, KNN and DT. This can be observed in cases 1, 2, and 3, where the meta-learner model gives better results than the base models, with an increase of 1.5% in all the cases. We have the following observation critical points from this analysis:

a. Optimal Split: The unique 70%,15%, and 15% split provides the balanced compromise between training and the capability to validate data and check efficiently, as in this case, it is observed that we have achieved higher performance metrics across all models.

b. Training Data vs. Evaluation Balance: Increasing training by 80%, 10%, and 10% slightly low model performance because of decreased assessment capability, at the same time as growing assessment statistics by 60%, 20%, and 20% limits training records an excessive amount of, negatively impacting version learning model.

c. Meta-Learner Robustness: The GBM Meta-Learner usually performs better across distinct splits, highlighting its functionality to combine and decorate base model predictions successfully.

4.2 Model Implementation Analysis

The model implemented as the base model is trained on the case1 ratio as explained in the Table 6, i.e. train 70%, validate 15% and test 15%. Taking this ratio as the standard ratio for the research, we have analyzed the follow-up for the base model and the meta-learner model. For both base models, KNN and DT without smote, the result is shown in the Table 7 and Figure 6 on the imbalance dataset. In contrast, the base model KNN and DT with smote results are shown in the Table 8 and Figure 7, as the observation DT outperforms as compared to the KNN with 88% on the unbalanced dataset, similarly with smote also DT performs better with the increase in accuracy of 1% as compared to KNN.

Table 6: Implementation Analysis Davide Chicco and Giuseppe Jurman Dataset

Case	Split Ratio	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Case 1	70% / 15% / 15%	KNN with SMOTE	0.87	0.86	0.87	0.85
		DT with SMOTE	0.88	0.89	0.88	0.88
		GBM Meta-Learner	0.92	0.92	0.89	0.86
Case 2	80% / 10% / 10%	KNN with SMOTE	0.86	0.84	0.83	0.81
		DT with SMOTE	0.88	0.87	0.85	0.89
		GBM Meta-Learner	0.89	0.88	0.88	0.88
Case 3	60% / 20% / 20%	KNN with SMOTE	0.81	0.81	0.81	0.81
		DT with SMOTE	0.77	0.75	0.79	0.77
		GBM Meta-Learner	82	81	83	82

Table 7: Base Model Implementation without Smote on Davide Chicco and Giuseppe Jurman Dataset

Model	Accuracy	Precision	Recall	F1-score
KNN	0.87	0.86	0.87	0.85
DT	0.88	0.89	0.88	0.88

Table 8: Base Model Implementation with Smote on Davide Chicco and Giuseppe Jurman Dataset

Model	Accuracy	Precision	Recall	F1-score
KNN	0.88	0.85	0.89	0.85
DT	0.9	0.9	0.9	0.9

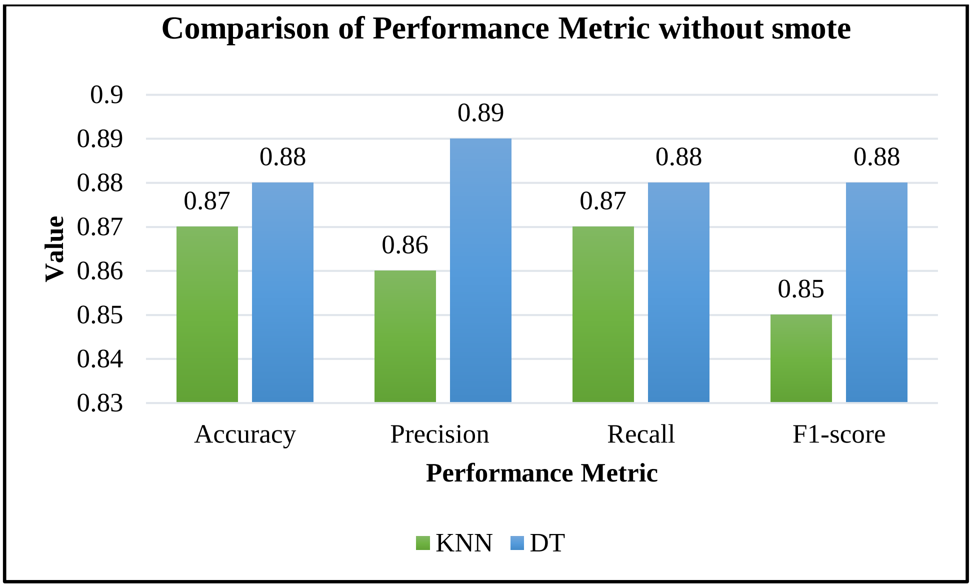

Figure 6: Comparison of Performance Metric without on Smote Davide Chicco and Giuseppe Jurman Dataset

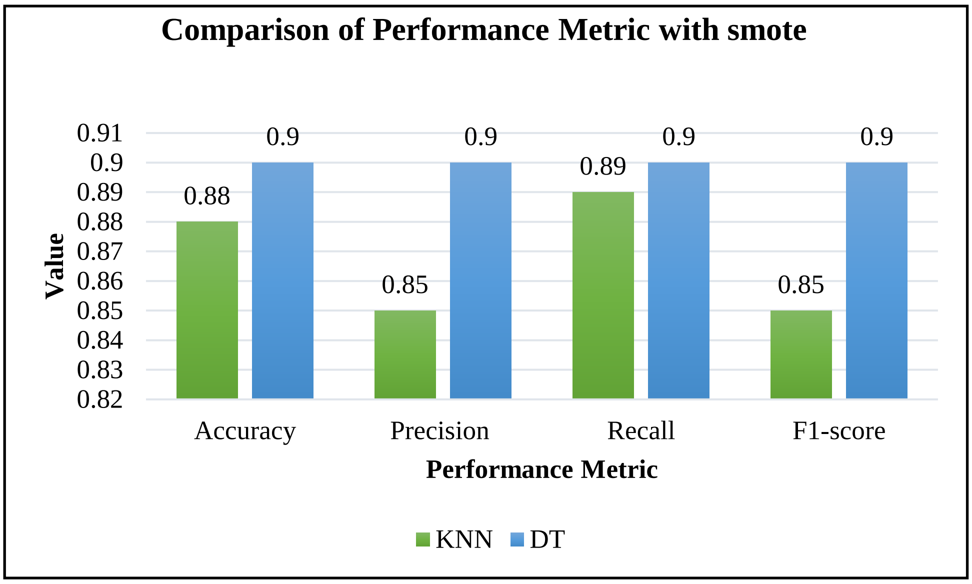

Figure 7: Comparison of Performance Metric without Smote on Davide Chicco and Giuseppe Jurman Dataset

The GBM is given the process of a feature from the validated dataset; hence, the model is not required to test on or without smote. The performance of the meta learner model GBM is explained in the Table 9 and Figure 8; as shown, the GBM meta learner performs well with an accuracy of 92%.

Table 9: Performance Analysis of GBM Meta Learner on Davide Chicco and Giuseppe Jurman Dataset

Model	Accuracy	Precision	Recall	F1-score
GBM	0.92	0.92	0.89	0.86

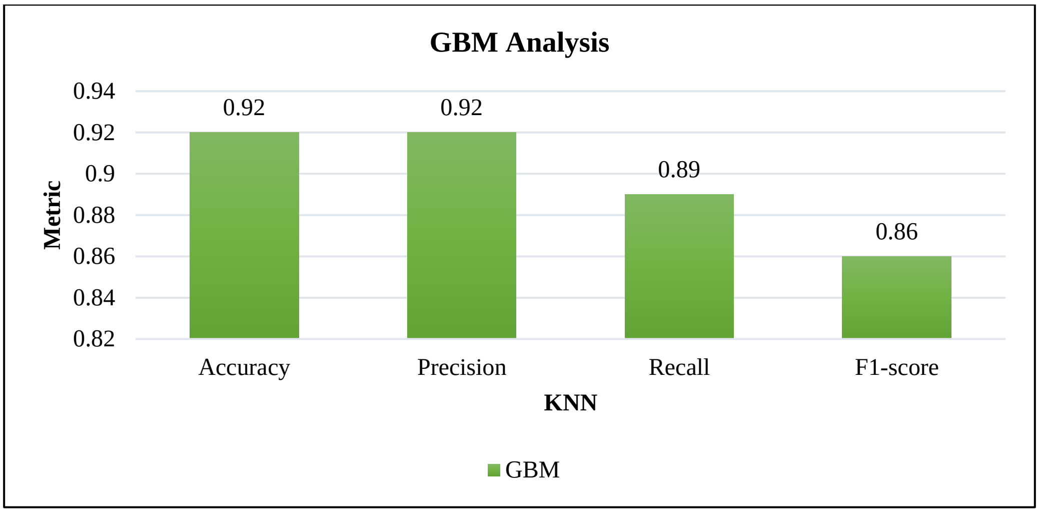

Figure 8: GBM Performance Analysis Davide Chicco and Giuseppe Jurman Dataset

5.3. Ideal Case Analysis

The research is a deep analysis of the implementation of the basic machine learning model, as the individual model cannot give the highest accuracy for heart disease prediction with this dataset. Hence, in the observation, we have analyzed the ideal case of the proposed model with the ratio of 70% for the train, 15% validate, and 15% test as per the results we have got for case1, for we have got the following observation as shown in the Table 10 and Figure 9 it indicates that all three models perform well. Still, the GBM outperforms the others, most likely owing to its ability to detect complicated patterns in data via learning, which combines several vulnerable novices to produce a robust prediction model. This shows that combination methods like GBM could provide a strategic advantage in forecasting outcomes more accurately for this dataset.

Table 10: Model Performance Analysis on Davide Chicco and Giuseppe Jurman Dataset

Model	Accuracy	Precision	Recall	F1-score
KNN Without Smote	0.87	0.86	0.87	0.85
DT Without Smote	0.88	0.89	0.88	0.88
GBM	0.92	0.92	0.89	0.86

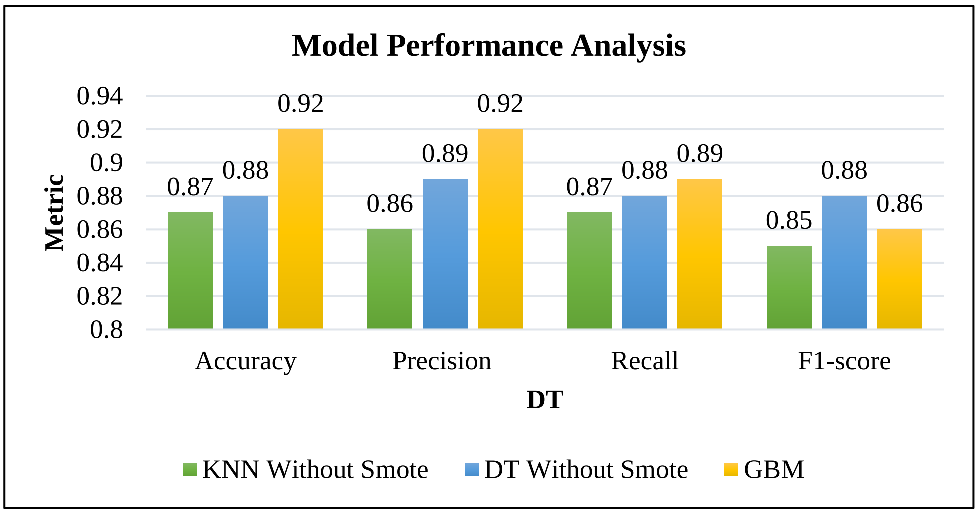

Figure 9: Model Performance Analysis on Davide Chicco and Giuseppe Jurman Dataset

5.4. Error Analysis

The graph in the Figure 10 illustrates the training process of a GBM model on the validated data, wherein the y-axis represents an error metric that the model is reduced over successive iterations (x-axis). The graph indicates that the minimum error is 0.2777777777777778 at K = 4. For the GBM with validated data, the Figure 11 error graph on the GBM model on the test data has a minimum error of 0.26666666666666666 at K = 24. Hence, we conclude that the GBM meta-learner model performs well on testing data with a reduced error rate

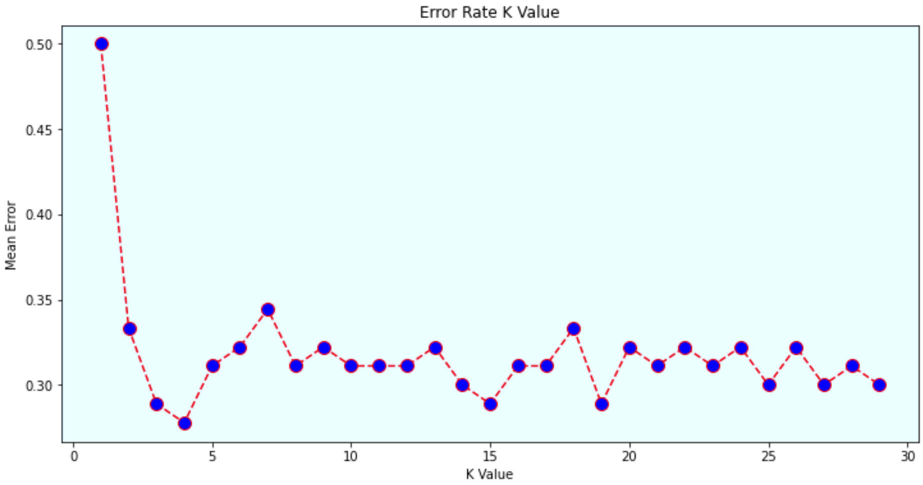

Figure 10: Error Rate K Value of GBM on Validate Data on Davide Chicco and Giuseppe Jurman Dataset

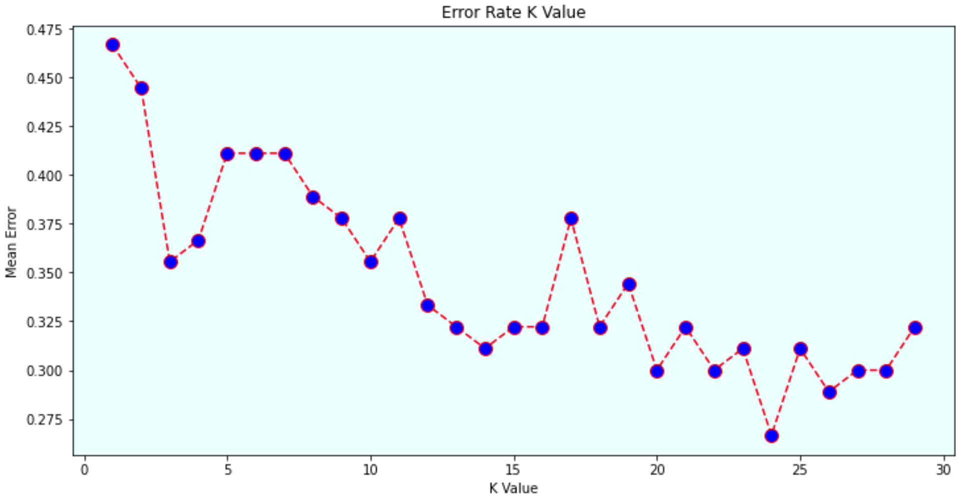

Figure 11: Error Rate K Value of GBM on Test Data Davide Chicco and Giuseppe Jurman Dataset

V. CONCLUSION

The research concludes the implementation of the proposed model and describes an in-depth comparison of the machine learning model for heart diseases. The performance was evaluated with base models KNN and DT on and without smote. Thus, the comparison stated the increase in accuracy with the DT. Still, the model did not perform as per the capacity in the other training cases, and the proposed model GBM as meta learner outperformed with an accuracy of 92% and precision of 92%, while recall of 0.89% and F1 score of 0.86%. Thus, the research explains the need for careful data partitioning to apply the various machine learning approaches with strategically smote and smote. Using the proposed work, the module has achieved the highest accuracy in predicting heart diseases using meta-learners. The future advancement for the work focuses on implementing the model on a larger dataset and analysing the cases according to the complex pattern of the machine learning models with images and series.

REFERENCES

[1] M. A. Bouqentar, O. Terrada, D. Lamrani, A. Ouhmida, B. Cherradi and A. Raihani, “Primary prediction of heart disease using machine learning algorithms and SMOTE,” *in* 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*,* Mohammedia, Morocco, pp. 1-7, 2023.

[2] Yang, Jian, and Jinhan Guan, “A Heart disease prediction model based on feature optimization and Smote-Xgboost Algorithm,” Information, vol.13, no.10, pp. 475, 2022.

[3] Wang K, Tian J, Zheng C, Yang H, Ren J, Li C, Han Q, Zhang Y, "Improving risk identification of adverse outcomes in chronic heart failure Using SMOTE+ENN and machine learning," Risk Manag Healthc Policy, vol.14, no.1, pp.2453-2463, 2021.

[4] Waqar, Muhammad, Hassan Dawood, Hussain Dawood, Nadeem Majeed, Ameen Banjar, and Riad Alharbey, "An efficient SMOTE-based deep learning model for heart attack prediction," Scientific Programming, vol.1, no.1, pp.1-12, 2021.

[5] Al Ahdal, Ahmed, Manik Rakhra, Rahul R. Rajendran, Farrukh Arslan, Moaiad Ahmad Khder, Binit Patel, Balaji Ramkumar Rajagopal, and Rituraj Jain, "Monitoring cardiovascular problems in heart patients using machine learning," Journal of healthcare engineering, vol.1, no.1, pp.1-10, 2023.

[6] Nashif, Shadman, Md Rakib Raihan, Md Rasedul Islam, and Mohammad Hasan Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," World Journal of Engineering and Technology, vol.6, no.4, pp.854-873, 2018.

[7] Patil, Shubham, Abhishek Yadav, Prof Akhtar Raza, and Parvez Rahi, "Heart disease prediction using machine Learning,", International Journal of Scientific Research in Science and Technology, vol.10, no.3, pp. 398-404, 2023.

[8] Srivastava, Asmit, and Ashish kumar Singh, "Heart disease prediction using machine learning," In 2nd IEEE International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 2633-2635, 2022.

[9] K., Jagadeesh., Raghavendran, R, "Heart disease prediction using machine learning," International Journal of Advanced Research in Science, Communication and Technology, vol.1, no.1, pp.408-415,2022.

[10] K. Battula, R. Durgadinesh, K. Suryapratap and G. Vinaykumar, "Use of machine learning techniques in the prediction of heart disease," 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Mauritius, Mauritius, pp. 1-5,2021.

[11] Lee, Eugene, Evan Chen, and Chen-Yi Lee, "Meta-rppg: remote heart rate estimation using a transductive meta-learner," In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, pp. 392-409. Springer International Publishing, 2020.

[12] M. Shuja, A. Qtaishat, H. M. Mishra, M. Kumar and B. Ahmed, "Machine learning to predict cardiovascular disease: systematic meta-analysis," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, pp. 1-6, 2023.

[13] I. Salem, R. Fathalla and M. Kholeif, "A Deep meta-learning framework for heart disease prediction," 2019 IEEE 15th International Scientific Conference on Informatics, Poprad, Slovakia, pp. 000483-000490, 2019.