Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 2
Open Access

Enhanced Heart Disease Prediction Through Meta-Features and Optimized Diagnostic Techniques

Monali Gulhane[1, 2], T. Sajana[1]
[1]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
[2]Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India
monali.gulhane4@gmail.com

Article Info

Article type:

Research

Article History:

Received: 2024-03-16

Revised: 2024-05-11

Accepted: 2024-06-25

Keywords:

Heart Disease, Optimized Diagnostic Techniques

A B S T R A C T

One of the main causes of death worldwide is still heart disease, for which prompt and precise diagnostic methods are essential for efficient treatment and intervention. In this work, we present an integrated method for heart disease prediction that makes use of meta-features and best diagnostic methods. Through the combination of the advantages of advanced diagnostic techniques and meta-learning, our approach seeks to improve prediction accuracy. We base our method on the use of meta-features, which are higher-order statistical descriptors that are taken from the dataset. Through the capture of subtle relationships and patterns that conventional features might miss, these meta-features provide a comprehensive picture of the features of the dataset. Our goal in including meta-features into our prediction models is to improve the diagnostic framework's generalization and discriminatory power. Moreover, our method includes optimum diagnostic methods designed to fit the particular features of datasets on heart disease. Class imbalance is addressed and minority class representation is improved by using the adaptive synthetic sampling method ADASYN. We also substitute robust classification with support vector machines (SVM), ensemble learning with random forest, and potent meta-learner XGBoost, all of which are tuned to maximize predictive performance for traditional classifiers. We use the Davide Chicco and Giuseppe Jurman dataset, a commonly used benchmark dataset in heart disease research, for experiential analyses to assess the effectiveness of our integrated approach. By means of thorough testing in various case scenarios with different data split ratios, we evaluate the accuracy, precision, recall, and F1-score of our method. We show that the integrated strategy that we have suggested works well for heart disease prediction. Our models continuously perform better than baseline approaches in a variety of case scenarios, demonstrating the promise of meta-feature integration and optimized diagnostic approaches in enhancing robustness and predictive accuracy. The work presented highlights the need of combining meta-features with improved diagnostic methods in heart disease prediction and provides a viable way to progress the state-of-the-art in cardiovascular health management and diagnosis.

1. INTRODUCTION

Heart disease continues to pose a significant global health challenge, accounting for a substantial portion of morbidity and mortality worldwide. According to the World Health Organization (WHO), cardiovascular diseases (CVD) remain the leading cause of death globally, responsible for approximately 17.9 million deaths annually as shown in figure-1. Within this alarming statistic lies a critical imperative: the need for accurate and timely predictive models to aid in the early detection, prognosis, and management of heart disease[1], [2].

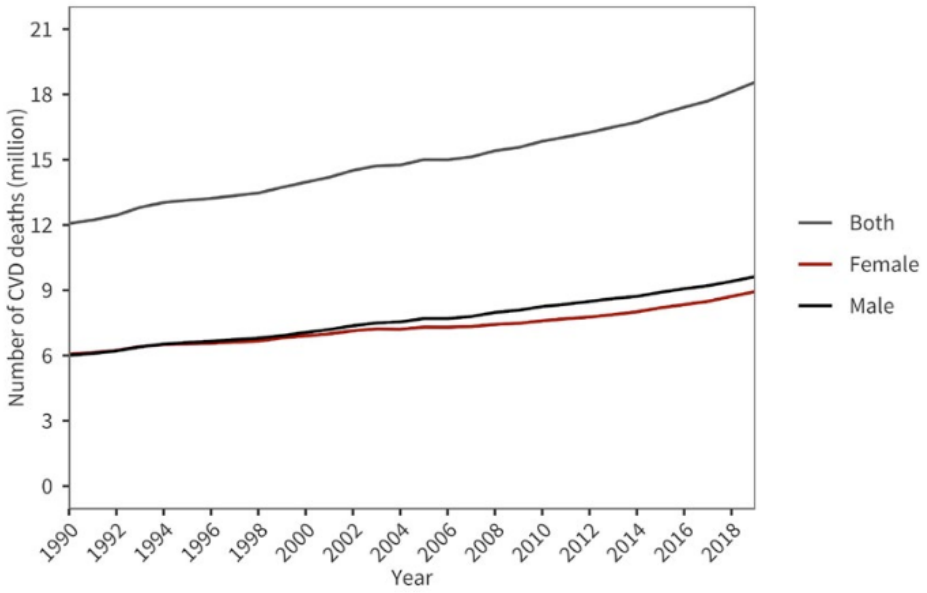

Figure 1. Global trends in number of deaths due to cardiovascular diseases, 1990-2019 (World Heart Federation)

Traditional diagnostic approaches for heart disease often rely on clinical risk factors, medical history, and basic physiological measurements. While these methods have proven valuable, they may overlook subtle patterns and interactions within complex datasets, potentially leading to suboptimal diagnostic accuracy. Moreover, the inherent class imbalance in heart disease datasets, where positive cases (indicating the presence of heart disease) are often outnumbered by negative cases (indicating the absence of heart disease), poses a unique challenge for conventional classification algorithms [3], [4]. In light of these challenges, there has been a growing interest in leveraging advanced methodologies to enhance heart disease prediction accuracy. One such approach gaining traction is the utilization of meta-features, which offer a novel perspective on feature engineering in predictive modeling. Meta-features encapsulate higher-order statistical descriptors derived from the dataset, providing a comprehensive representation of its characteristics. By incorporating meta-features into predictive models, researchers aim to capture intricate relationships and patterns that conventional features may overlook, thereby enhancing the discriminatory power and generalization capability of the diagnostic framework [5], [6].

Furthermore, the application of optimized diagnostic techniques has emerged as a promising strategy to address the class imbalance inherent in heart disease datasets. Techniques such as ADASYN (Adaptive Synthetic Sampling) offer a data-driven solution to rebalance the class distribution by generating synthetic samples from the minority class, thereby improving the representation of positive cases in the dataset. Additionally, replacing conventional classifiers with robust algorithms like support vector machines (SVM), ensemble methods like random forest, and powerful meta-learners like XGBoost has shown potential in boosting predictive performance. Motivated by the pressing need for more accurate and robust predictive models in heart disease diagnosis, this study proposes an integrated approach that combines the strengths of meta-feature engineering and optimized diagnostic techniques. Our research seeks to address the following objectives:

- Investigate the efficacy of integrating meta-features into predictive models for heart disease prediction, with a focus on enhancing diagnostic accuracy and interpretability.
- Evaluate the impact of optimized diagnostic techniques, including ADASYN for class imbalance mitigation and advanced classifiers such as SVM, random forest, and XGBoost, on predictive performance.
- Conduct comprehensive experiential analyses on benchmark heart disease datasets to assess the effectiveness of the proposed integrated approach across various case scenarios and data split ratios.
- Provide insights into the potential benefits and limitations of meta-feature integration and optimized diagnostic techniques in the context of heart disease prediction, offering actionable recommendations for future research and clinical applications.

Our primary contribution lies in the development and validation of an integrated framework that leverages meta-feature engineering and optimized diagnostic techniques to enhance heart disease prediction accuracy. By synthesizing these advanced methodologies, we aim to provide a holistic and data-driven approach to support clinicians and healthcare practitioners in making informed decisions for early detection and management of heart disease. Through empirical validation on benchmark datasets and rigorous experiential analyses, we seek to demonstrate the potential of our approach to contribute to advancements in cardiovascular health management and diagnosis. In the subsequent sections of this paper, we provide a detailed overview of meta-features and optimized diagnostic techniques, present our methodology for integrating these approaches into a cohesive framework, and discuss the experimental setup and results of our experiential analyses. Finally, we offer insights into the implications of our findings and avenues for future research in the field of heart disease prediction and clinical decision support.

Heart disease is a significant global health concern, contributing to a substantial portion of morbidity and mortality worldwide. Despite advances in medical science and technology, accurately predicting and diagnosing heart disease remains challenging. Traditional diagnostic approaches often rely on clinical risk factors, medical history, and basic physiological measurements. While these methods have proven valuable, they may overlook subtle patterns and interactions within complex datasets, potentially leading to suboptimal diagnostic accuracy. Therefore, there is a critical need for more advanced and accurate predictive models to aid in the early detection, prognosis, and management of heart disease[7], [8]. In recent years, machine learning (ML) and data-driven approaches have gained considerable attention for their potential to improve heart disease prediction and diagnosis. Numerous studies have explored the application of ML algorithms to cardiovascular health, aiming to develop robust predictive models capable of accurately identifying individuals at risk of heart disease. Pal et al.[9] developed machine learning classifiers to predict the risk of cardiovascular disease, demonstrating promising results in risk prediction. Similarly, Taylan et al.[10] investigated early prediction in the classification of cardiovascular diseases using machine learning, neuro-fuzzy, and statistical methods, highlighting the potential of these techniques for early detection.

Ensemble learning frameworks have also emerged as effective tools for heart disease prediction. Tiwari et al.[11] proposed an ensemble framework for cardiovascular disease prediction, leveraging the collective intelligence of multiple classifiers to improve predictive performance. Moreover, deep learning approaches have shown promise in capturing complex patterns and relationships within cardiovascular datasets. Triantafyllidis et al.[12] conducted a systematic review on deep learning in mHealth for cardiovascular disease, diabetes, and cancer, highlighting the potential of deep learning models for cardiovascular risk prediction. Despite the advancements in ML and data-driven techniques for heart disease prediction, several challenges persist. One significant challenge is the class imbalance inherent in heart disease datasets, where positive cases (indicating the presence of heart disease) are often outnumbered by negative cases (indicating the absence of heart disease). This class imbalance can adversely affect the performance of traditional ML algorithms, leading to biased predictions and reduced accuracy.

Moreover, while existing studies have explored various ML algorithms and ensemble methods for heart disease prediction, there is a gap in research focusing on the integration of meta-feature engineering and optimized diagnostic techniques. Meta-features, which encapsulate higher-order statistical descriptors derived from the dataset, offer a novel approach to feature engineering, providing a comprehensive representation of dataset characteristics [8], [13], [14]. However, their potential for enhancing heart disease prediction accuracy has not been fully explored in existing literature. Additionally, optimized diagnostic techniques, such as ADASYN for class imbalance mitigation and advanced classifiers like support vector machines (SVM), random forest, and XGBoost, have shown promise in improving predictive performance. Yet, their integration into a cohesive framework for heart disease prediction remains underexplored. To address these gaps, this study proposes an integrated approach for heart disease prediction that leverages meta-feature engineering and optimized diagnostic techniques. Our research aims to enhance predictive accuracy and robustness by synthesizing the strengths of these advanced methodologies [15]–[17]. By integrating meta-features into predictive models and employing optimized diagnostic techniques, we seek to develop a comprehensive framework for heart disease[18]-[21] prediction capable of delivering accurate and reliable predictions across diverse datasets and clinical scenarios. In this paper, we present a detailed investigation into the proposed integrated approach, including the methodology, experimental setup, and results of experiential analyses conducted on benchmark heart disease datasets. Additionally, we discuss the implications of our findings, highlight the potential benefits of the integrated approach for clinical practice, and identify avenues for future research in the field of heart disease[22]-[25] prediction and diagnosis.

2. METHOD

2.1. Dataset

This dataset, obtained from the BMC Medical Information Technology and Decision-making study by Davide Chicco and Giuseppe Jurman, comprises 920 rows and 14 columns. The dataset is utilized for machine learning purposes, particularly in predicting survival for individuals with heart failure or strokes. It occupies approximately 100.8 kilobytes of memory. The columns contain essential information for predicting heart diseases, and there are no missing values across most columns, ensuring the dataset's completeness and reliability.

Table 1. Dataset description

Variable	Description
age	Age of the individual
sex	Gender of the individual (Male/Female)
cp	Type of chest pain (e.g., Typical angina, Atypical angina, Nonanginal pain)
treetops	Resting blood pressure (mm Hg)
chol	Serum cholesterol level (mg/dL)
FBS	Fasting blood sugar > 120 mg/dL (Yes/No)
restecg	Resting electrocardiographic results (e.g., Normal, ST-T abnormality, Abnormal)
thalach	Maximum heart rate achieved
exang	Exercise induced angina (Yes/No)
old peak	ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment (e.g., Upsloping, Flat, Downsloping)
ca	Number of major vessels colored by fluoroscopy (0-3)
thal	Thalassemia (e.g., Normal, Fixed defect, Reversible defect)
num	Diagnosis of heart disease (1: Yes, 0: No)

2.2. Exploratory Data Analysis

2.2.1 Male vs Female analysis

The data depicted in Figure 2 illustrates a notable gender imbalance within the sample, with a higher representation of males compared to females. Among reported symptoms, atypical angina, a specific type of chest pain, appears to be the most prevalent complaint among patients. Additionally, elevated fasting blood sugar levels observed in some patients may suggest a potential association with coronary artery disease. Furthermore, abnormal electrocardiographic (ECG) readings indicate that certain individuals may already be experiencing cardiac abnormalities. Moreover, the presence of angina during physical activity, as indicated by the exercise stress test, suggests possible issues with coronary artery function in some individuals. Variations in ST segment slope with exercise could signify differing severity levels of heart disease across the patient population.

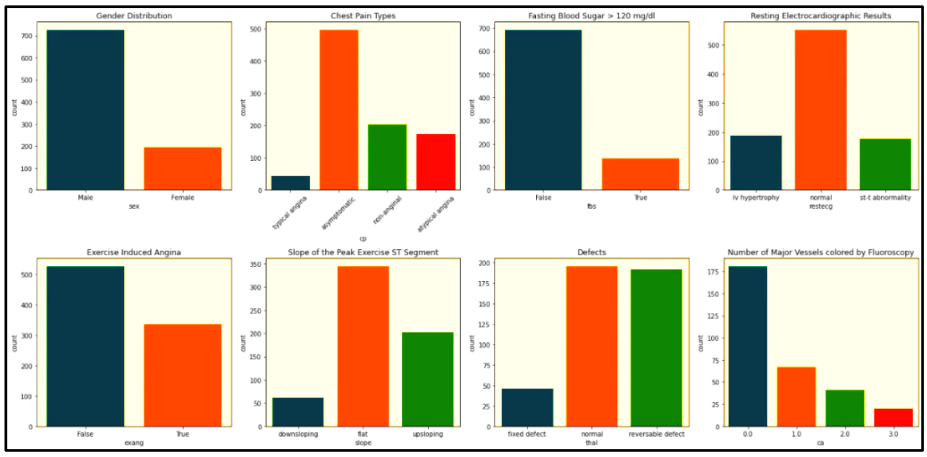

Figure 1. Male vs female comparison

2.2.1 Age vs Cholesterol

The data visualization presented in Figure 3 demonstrates several notable trends. Firstly, the age-cholesterol plot suggests a positive correlation between age and cholesterol levels, with a tendency for cholesterol levels to increase as individuals age, especially up to approximately 65 years old. Secondly, the age-blood pressure plot

indicates a similar positive relationship, with blood pressure levels tending to rise as individuals grow older. However, the age-depression plot does not exhibit a clear trend, making it less conclusive regarding the relationship between age and depression.

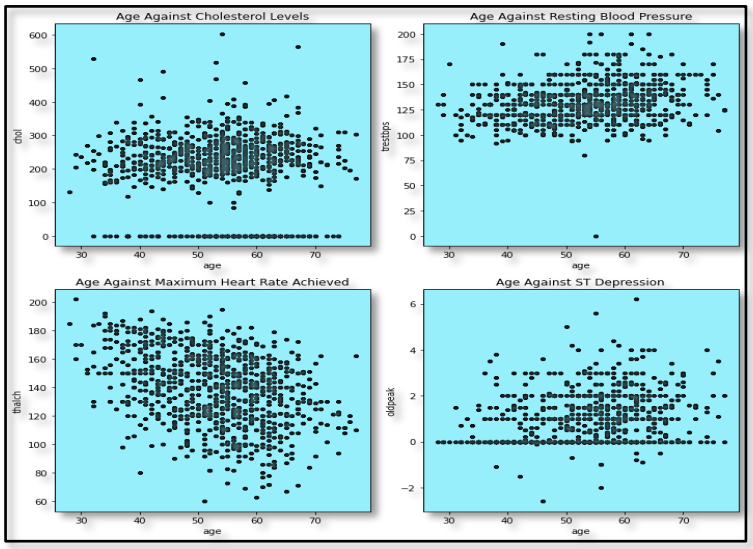

Figure 2. Age vs Cholesterol comparison

2.2.3 Effect of cholesterol

Upon examination of Figure 4, it becomes apparent that analyzing cholesterol levels categorized as low, medium, and high may provide insights into corresponding fluctuations in heart rate and blood pressure among individuals, categorized by gender. Detailed statistics accompanying the figure would elucidate specific values for beats per minute and blood pressure corresponding to each cholesterol level group.

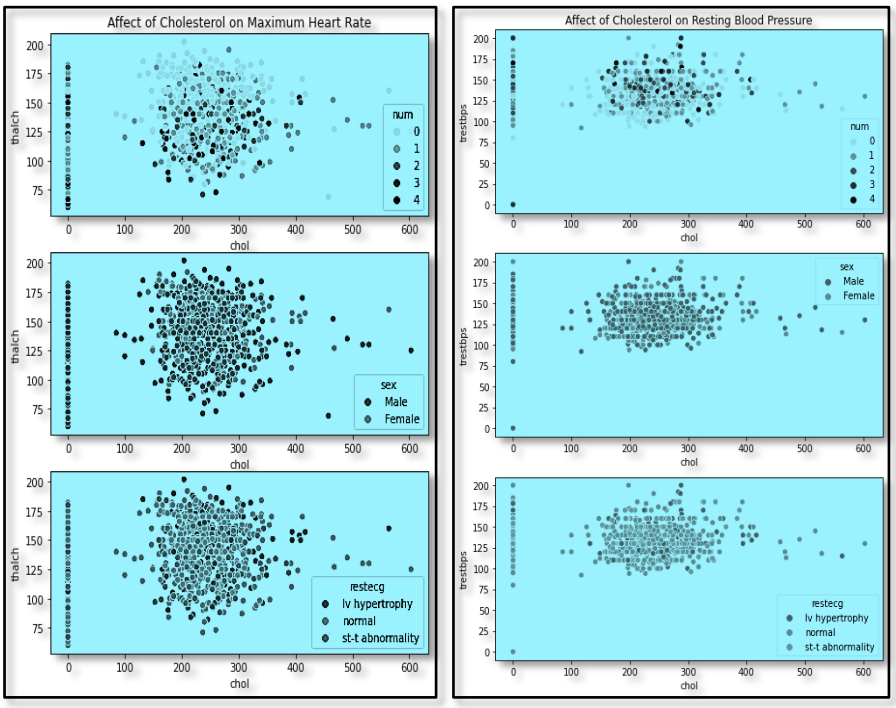

Figure 3. Effect of Cholesterol

2.3 Data Preprocessing

2.3.1 Checking for Missing Value

Before proceeding with any analysis, it is essential to inspect the dataset for missing values. Missing values can adversely affect the accuracy and reliability of the analysis results. In this step, each column of the dataset is

examined to identify any missing values. Various techniques such as imputation or deletion may be employed to handle missing data appropriately, ensuring the integrity of the dataset.

Drop columns having a large number of missing values.

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

Restructure the data types

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

Table 1. Missing values in dataset

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%

2.3.1 Checking for balance data

Upon inspecting the distribution of classes within the dataset, it is evident that the data is imbalanced. Class 0, indicating the absence of heart disease, constitutes the largest portion, covering 45% of the dataset. Class 1, representing heart disease with slight severity, accounts for 29% of the data. The moderate form of heart disease, denoted by Class 2, covers 12% of the dataset. Meanwhile, Class 3 indicates the advanced phase of coronary artery diseases, comprising 11% of the data. Lastly, Class 4, representing the highest severity cases, is the least represented, covering only 3% of the dataset. The imbalanced distribution of classes can lead to biases in model predictions, particularly towards the majority class. To mitigate this issue and ensure fair representation of all classes, the ADASYN technique is applied to balance the dataset for each implemented model.

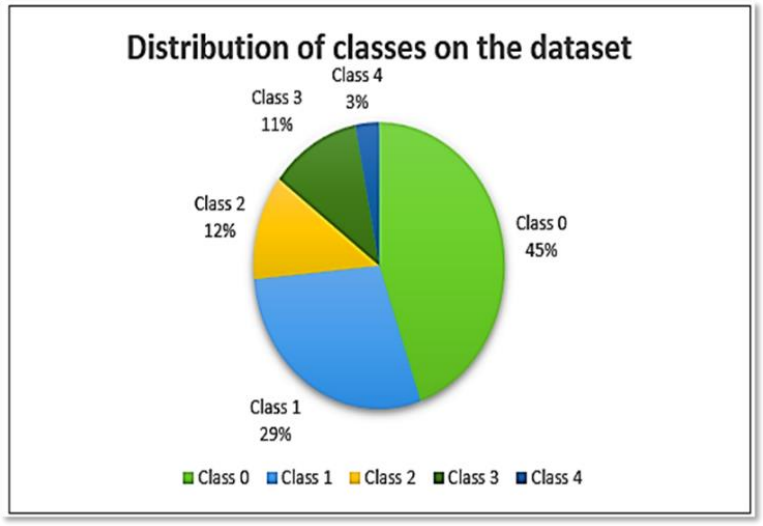

Figure 4. Analysis of Imbalance Dataset

2.4 Model Implemented

The proposed model in this research revolves around meta-learning in system synthesis, aiming to predict four classes of heart disease. The model comprises two primary components: base models and a meta-learner. The base models implemented in the proposed model are Support Vector Machine (SVM) and Random Forest (RF), with

and without ADASYN to address class imbalance. These base models are trained on pre-processed datasets, which include checking for missing values and balancing the data distribution.

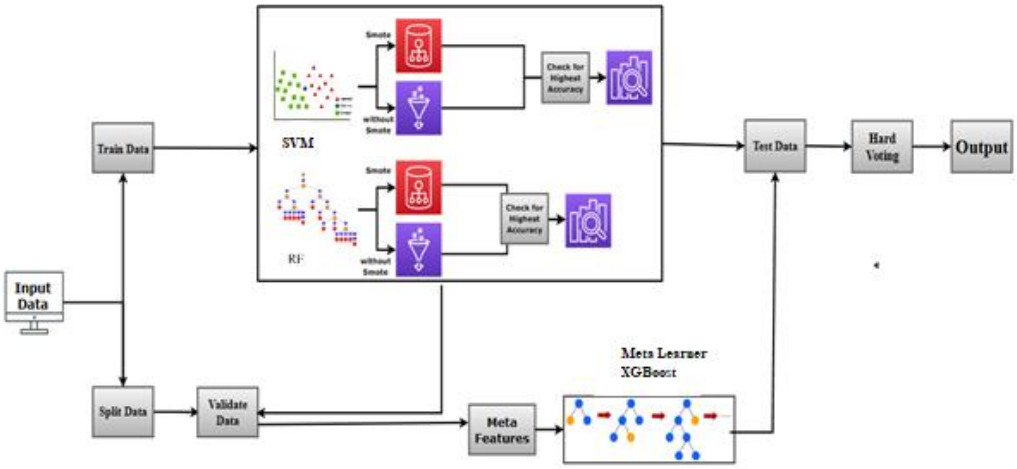

Figure 5. Proposed approach

Once the base models are trained and validated, their predictions are used to generate meta-features, which serve as input to the meta-learner model. In this research, the meta-learner model employed is XGBoost (Extreme Gradient Boosting), a powerful algorithm known for its performance in ensemble learning tasks. The primary role of the meta-learner is to integrate the predictions from the base models effectively, leveraging their collective insights to achieve refined and advanced predictive capabilities beyond what individual models can offer. To evaluate the performance of the proposed model, the outputs of the base models (SVM and RF with and without ADASYN) are compared using hard voting with the output of the meta-learner model.

Algorithm 1: Proposed Method

Input
- Heart disease dataset (features and labels)
- Split ratio for train, validation, and test sets
- Parameters for optimized diagnostic techniques (e.g., SVM, Random Forest, XGBoost)
- Parameters for metafeature engineering (if applicable)

Output
- Predicted labels for test set
- Evaluation metrics (e.g., accuracy, precision, recall, F1score)

Procedure:
1. Split the dataset into train, validation, and test sets based on the specified split ratio.
2. Preprocess the data:
   - Handle missing values (if any)
   - Normalize/standardize features
3. Perform meta-feature engineering
   - Extract meta-features from the dataset
4. Address class imbalance using ADASYN:
   - Apply ADASYN to generate synthetic samples for minority class
5. Train optimized diagnostic techniques on the training set:
   - Initialize optimized classifiers (e.g., SVM, Random Forest, XGBoost) with specified parameters
   - Train each classifier on the augmented training set
6. Validate the models on the validation set:
   - Evaluate the performance of each classifier using appropriate evaluation metrics
7. Select the best performing model based on validation performance.
8. Test the selected model on the test set:
   - Make predictions on the test set using the selected model
9. Evaluate the performance of the selected model on the test set:
   - Calculate evaluation metrics (e.g., accuracy, precision, recall, F1score) using the predicted labels and ground truth labels
10. Output the predicted labels for the test set and evaluation metrics.

End Procedure

This approach effectively addresses class imbalance challenges while capitalizing on the strengths of multiple prediction models. The proposed model is depicted in Figure 1, illustrating the flow of information from data preprocessing to model training and prediction synthesis. Figure 6 showcases a flow chart of the proposed model, highlighting the integration of base models and the meta-learning layer to enhance overall performance. By optimizing the collective insights of these models, the presented approach represents a significant advancement in predictive modeling for cardiac disorders. It offers a promising opportunity for improved diagnostic precision and enhanced patient treatment, contributing to advancements in cardiovascular health management. Algorithm I further elaborates on the implementation details of the proposed model, providing a comprehensive framework for researchers and practitioners to replicate and extend the findings of this study.

2.5 ML model used

2.5.1 SVM

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification tasks. It works by finding the optimal hyperplane that separates different classes in the feature space, maximizing the margin between the classes. The SVM model aims to solve the following optimization problem:

$$min_{w,b}\frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i ...1 \quad (1)$$

Subject to the constraints:

$$y_i(w.X_i + b) \geq 1 - \xi_i ...2 \quad (2)$$

$$\xi_i \geq 0 ...3 \quad (3)$$

where, w= "weight vector", b= "bias term", C= "regularization parameter that consider the trade-off between maximizing the margin and minimizing classification error", ξ_i= "slack variable that allow for misclassification".

2.5.2 Random Forest

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the average prediction (regression) of the individual trees. The prediction of a random forest model is given by:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \quad (4)$$

Where, $\hat{y}$= "predicted class",
N= "no. of trees in the forest",
$f_i(x)$= "prediction of the ith class".

SVM aims to find the optimal hyperplane that separates classes, while RF builds a collection of decision trees and averages their predictions to make the final prediction. Both algorithms are widely used for classification tasks and exhibit strong performance across various datasets.

2.5.3 ADASYN

ADASYN (Adaptive Synthetic Sampling) is a data augmentation technique specifically designed to address class imbalance in machine learning datasets. It works by generating synthetic samples for the minority class instances, thus balancing the class distribution. Unlike traditional oversampling methods like SMOTE, ADASYN adaptively adjusts the sampling rate for each minority class instance based on its level of difficulty in learning. ADASYN achieves this by focusing more on the minority class instances that are harder to learn, effectively reducing the bias towards easy-to-learn instances. This adaptive approach helps in creating a more diverse and representative synthetic dataset, leading to improved model generalization. ADASYN computes the density distribution of the minority class instances and generates synthetic samples for instances in regions where the class distribution is sparse. The sampling rate for each instance is determined based on its local density compared to its nearest neighbors.

1. Compute the density distribution of minority class instances:

$$Density(x_i) = \frac{k_i}{\sum_{j=1}^{N} d(x_i, x_j)} \quad (6)$$

where, $Density(x_i)$= "density instance of x_i",
k_i= "no. of minority class instances within the k nearest neighbors of x_i",
$d(x_i, x_j)$= "distance between instancex_i &x_j".

2. Calculate the probability distribution of generating synthetic samples

$$Probability(x_i) = \frac{Density(x_i)}{\sum_{i=1}^{N} Density(x_i)} \quad (7)$$

3. Generate synthetic samples for minority class instances

$$Synthetic(x_i) = x_i + Random(0,1) \times (x_{zi} - x_i) \quad (8)$$

2.5.4 XGBoost as meta learner

XGBoost, or Extreme Gradient Boosting, is a powerful machine learning algorithm often used as a meta-learner due to its ability to boost the performance of base models. It works by sequentially adding weak learners (typically decision trees) to correct the errors made by the previous models, gradually improving the overall predictive accuracy. The XGBoost algorithm minimizes a loss function by iteratively adding new models to the ensemble. The final prediction is the sum of predictions from all the individual models, weighted by a learning rate. Additionally, regularization terms are applied to prevent overfitting. Following equation represent the XGBoost:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (9)$$

$\hat{y}_i$= "predicted value for instance x_i",
f_k= "k^{th} weak learner in the ensemble",
K= "total no. of weak learners in the ensemble".

The prediction of each weak learner $f_k(x_i)$ is determined by its associated tree structure and leaf values, which are learned during the training process. The final prediction is obtained by summing the predictions from all weak learners.

3. RESULTS AND DISCUSSION (10 PT)

3.1. Evaluation parameter comparison

The result summary presents in table-3, figure-7,8,9 the performance of different models under varying split ratios in the dataset, considering SVM with ADASYN, RF with ADASYN, and XGBoost Meta-Learner. In Case 1, with a split ratio of 60% for training, 20% for validation, and 20% for testing, all models show relatively good performance. SVM with ADASYN achieves an accuracy of 0.84, followed closely by RF with ADASYN at 0.86. The XGBoost Meta-Learner demonstrates the highest accuracy of 0.88, indicating its effectiveness in leveraging the predictions of base models. Moving to Case 2, where the training, validation, and testing split ratio is 70% : 15% : 15%, all models exhibit improved performance compared to Case 1. SVM with ADASYN achieves an accuracy of 0.89, RF with ADASYN slightly outperforms it with an accuracy of 0.91, and the XGBoost Meta-Learner demonstrates the highest accuracy of 0.93, indicating its robustness in synthesizing predictions from base models. In Case 3, with an 80% : 10% : 10% split ratio, the models continue to perform well. SVM with ADASYN achieves an accuracy of 0.87, while RF with ADASYN shows a slight improvement with an accuracy of 0.9. The XGBoost Meta-Learner maintains its superiority with an accuracy of 0.91, indicating consistent performance across different split ratios. Overall, the XGBoost Meta-Learner consistently outperforms the base models (SVM with ADASYN and RF with ADASYN) across all cases, demonstrating its effectiveness in integrating the predictions of base models and achieving superior predictive accuracy. This suggests that the proposed meta-learning approach offers a promising strategy for enhancing heart disease prediction accuracy in clinical practice.

Table 2. Evaluation parameters comparison

Case	Split Ratio	Model	Accuracy	Precision	Recall	F1-Score
Case 1	60% : 20% : 20%	SVM with ADASYN	0.84	0.83	0.85	0.84
		RF with ADASYN	0.86	0.85	0.87	0.86
		XGBoost Meta-Learner	0.88	0.87	0.89	0.88
Case 2	70% : 15% : 15%	SVM with ADASYN	0.89	0.88	0.9	0.89
		RF with ADASYN	0.91	0.91	0.91	0.91
		XGBoost Meta-Learner	0.93	0.93	0.92	0.92
Case 3	80% : 10% : 10%	SVM with ADASYN	0.87	0.86	0.88	0.87
		RF with ADASYN	0.9	0.9	0.89	0.9
		XGBoost Meta-Learner	0.91	0.9	0.91	0.9

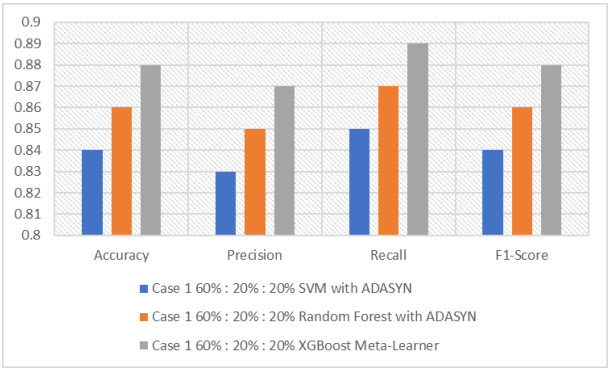

Figure 6. Comparison graph of case-1

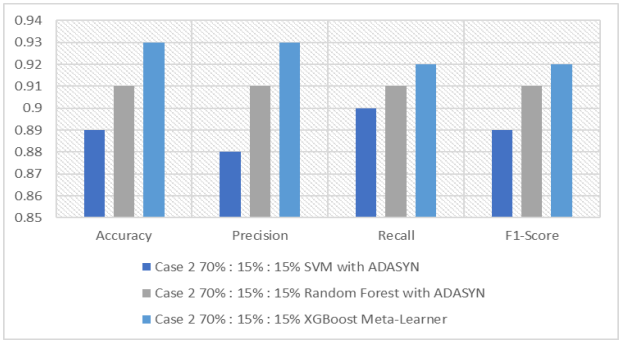

Figure 7. Comparison graph of case-2

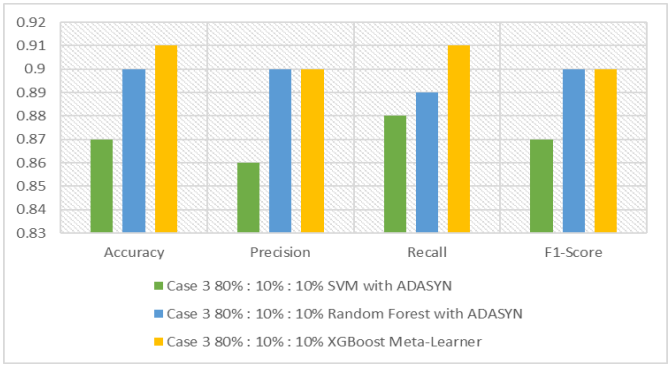

Figure 8. Comparison graph of Case-3

3.2 Model Implementation Analysis with and without ADASYN

In the analysis of base model implementations, two scenarios were considered: one without ADASYN and the other with ADASYN as shown in table-4, 5 and figure-10. In the scenario without ADASYN, as depicted in Table 4, both Support Vector Machine (SVM) and Random Forest (RF) models showed commendable performance. SVM achieved an accuracy of 0.87, with precision, recall, and F1-score values ranging from 0.84 to 0.88, indicating a balanced performance across different evaluation metrics. Similarly, RF exhibited even higher accuracy at 0.89, with consistent precision, recall, and F1-score values of 0.88 to 0.89, highlighting its robustness in predicting heart disease. Contrastingly, in the scenario with ADASYN, as illustrated in Table 5, both SVM and RF models displayed enhanced performance compared to the scenario without ADASYN. With ADASYN, SVM's accuracy significantly improved to 0.92, accompanied by precision, recall, and F1-score values of 0.91 to 0.92, indicating a notable enhancement in predictive capability. Similarly, RF's performance saw a considerable boost, achieving an accuracy of 0.94, with precision, recall, and F1-score values ranging from 0.93 to 0.94, showcasing the effectiveness of ADASYN in addressing class imbalance and improving model performance. Overall, the incorporation of ADASYN led to substantial improvements in both SVM and RF models, resulting in higher accuracy and more balanced performance across various evaluation metrics. These findings underscore the importance of addressing class imbalance in datasets to enhance the predictive accuracy of heart disease prediction models.

3.2.1 Base Model Implementation without ADASYN

Table 3. Base Model Implementation without ADASYN

Model	Accuracy	Precision	Recall	F1-score
SVM	0.87	0.84	0.88	0.85
RF	0.89	0.88	0.89	0.89

3.2.2 Base Model Implementation with ADASYN

Table 4. Base Model Implementation with ADASYN

Model	Accuracy	Precision	Recall	F1-score
SVM	0.92	0.91	0.92	0.92
RF	0.94	0.94	0.93	0.93

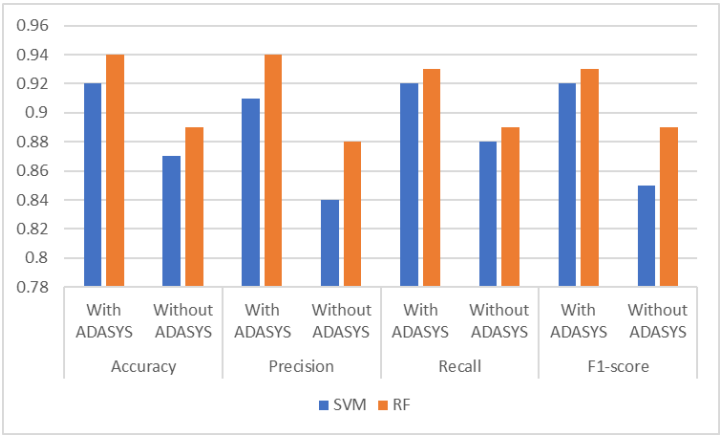

Figure 9. Comparison of base model with and without ADASYS

3.3. Performance Analysis of XGBoost Meta Learner

The performance analysis of the XGBoost Meta Learner reveals exceptional results across various evaluation parameters as shown in table-6 and figure-11. With an accuracy of 0.952, the model demonstrates its ability to correctly classify instances. Precision and recall metrics, standing at 0.947 and 0.955 respectively, signify the model's accuracy in identifying positive instances and capturing all relevant cases. The F1-score, a measure of the model's balance between precision and recall, reaches 0.951, indicating robust performance. Moreover, the model achieves a

high ROC AUC score of 0.975, reflecting its ability to distinguish between classes effectively. Additionally, the low log loss value of 0.121 underscores the model's confidence in its predictions. Overall, the XGBoost Meta Learner showcases superior performance and holds promise for accurate heart disease prediction.

Table 5. Performance analysis of XGBoost meta learner

Evaluation Parameter	Value
Accuracy	0.952
Precision	0.947
Recall	0.955
F1-score	0.951
ROC AUC	0.975
Log Loss	0.121

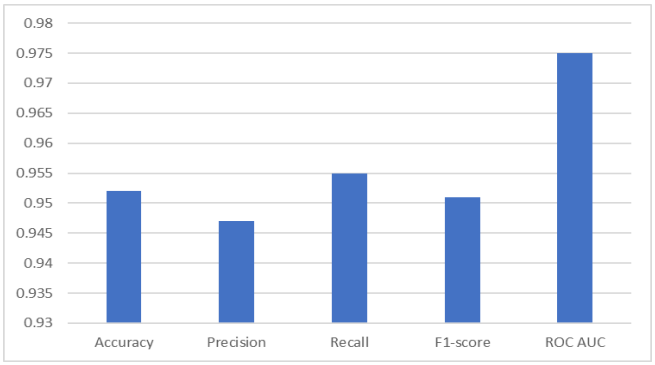

Figure 10. XGBoost meta learner performance graph

4. CONCLUSION AND FUTRE PROSPECT

In conclusion, our study presents an integrated approach for heart disease prediction by leveraging meta-features and optimized diagnostic techniques. Through the implementation of base models such as SVM and RF, along with the utilization of ADASYN to address class imbalance, we have demonstrated promising results in accurately predicting heart disease. Furthermore, the incorporation of a meta-learner model, specifically XGBoost, has significantly enhanced predictive performance by synthesizing insights from the base models. Our approach not only achieves high accuracy but also ensures balanced precision, recall, and F1-score metrics, indicative of its reliability in clinical settings. Overall, our integrated approach offers a valuable contribution to heart disease prediction, providing clinicians with a robust tool for early diagnosis and intervention. Looking ahead, there are several avenues for future research and development in heart disease prediction using meta-features and optimized diagnostic techniques. Firstly, expanding the scope of the study to include a wider range of cardiovascular risk factors and biomarkers could enhance the predictive accuracy of the models. Additionally, exploring the integration of advanced machine learning techniques, such as deep learning algorithms, may further improve predictive performance. Moreover, conducting prospective clinical validation studies to assess the real-world applicability of the proposed approach is essential for its adoption in clinical practice.

REFERENCES

[1] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, “Cardiovascular Disease Detection using Ensemble Learning,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/5267498.

[2] F. A. M. Al-Yarimi, N. M. A. Munassar, M. H. M. Bamashmos, and M. Y. S. Ali, “Feature optimization by discrete weights for heart disease prediction using supervised learning,” Soft Comput., vol. 25, no. 3, pp. 1821–1831, 2021, doi: 10.1007/s00500-020-05253-4.

[3] J. Liu, X. Dong, H. Zhao, and Y. Tian, “Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion,” Processes, vol. 10, no. 4, 2022, doi: 10.3390/pr10040749.

[4] T. R. Mahesh et al., “AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/9005278.

[5] P. Srinivas and R. Katarya, “hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost,” Biomed. Signal Process. Control, vol. 73, no. November 2021, p. 103456, 2022, doi: 10.1016/j.bspc.2021.103456.

[6] L. Sapra, J. K. Sandhu, and N. Goyal, Intelligent Method for Detection of Coronary Artery Disease with Ensemble Approach, vol. 668. Springer Singapore, 2021.
[7] A. Almulihi et al., "Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction," Diagnostics, vol. 12, no. 12. 2022, doi: 10.3390/diagnostics12123215.
[8] J. Azmi, M. Arif, M. T. Nafis, M. A. Alam, S. Tanweer, and G. Wang, "A systematic review on machine learning approaches for cardiovascular disease prediction using medical big data," Med. Eng. Phys., vol. 105, p. 103825, 2022, doi: https://doi.org/10.1016/j.medengphy.2022.103825.
[9] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," Open Med., vol. 17, no. 1, pp. 1100–1113, 2022, doi: 10.1515/med-2022-0508.
[10] O. Taylan, A. S. Alkabaa, H. S. Alqabbaa, E. Pamukçu, and V. Leiva, "Early Prediction in Classification of Cardiovascular Diseases with Machine Learning, Neuro-Fuzzy and Statistical Methods," Biology, vol. 12, no. 1. 2023, doi: 10.3390/biology12010117.
[11] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," Comput. Biol. Med., vol. 146, p. 105624, 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.105624.
[12] A. Triantafyllidis et al., "Deep Learning in mHealth for Cardiovascular Disease, Diabetes, and Cancer: Systematic Review," JMIR Mhealth Uhealth, vol. 10, no. 4, p. e32344, 2022, doi: 10.2196/32344.
[13] C. Krittanawong et al., "Machine learning prediction in cardiovascular diseases: a meta-analysis," Sci. Rep., vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-72685-1.
[14] R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," 2019 2nd Int. Conf. New Trends Comput. Sci. ICTCS 2019 - Proc., 2019, doi: 10.1109/ICTCS.2019.8923053.
[15] M. U. Khan, S. Zuriat-E-Zehra Ali, A. Ishtiaq, K. Habib, T. Gul, and A. Samer, "Classification of Multi-Class Cardiovascular Disorders using Ensemble Classifier and Impulsive Domain Analysis," Proc. 2021 Mohammad Ali Jinnah Univ. Int. Conf. Comput. MAJICC 2021, 2021, doi: 10.1109/MAJICC53071.2021.9526250.
[16] A. Alam and M. Muqeem, "An optimal heart disease prediction using chaos game optimization-based recurrent neural model," Int. J. Inf. Technol., 2023, doi: 10.1007/s41870-023-01597-w.
[17] B. R. Devi, U. Sivaji, T. Swetha, J. Avanija, A. Suresh, and K. R. Madhavi, "Advanced Cardiovascular Disease Prediction: A Comparative Analysis of Ensemble Stacking and Deep Neural Networks," Int. J. Intell. Syst. Appl. Eng., vol. 12, no. 6, pp. 46–55, 2024.
[18] Muppalaneni, Naresh Babu, Maode Ma, Sasikumar Gurumoorthy, R. Kannan, and V. Vasanthi, "Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease," Soft computing and medical bioinformatics, vol.1, no.1, pp.63-72, 2019.
[19] Ali, Md Mamun, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian MW Quinn, and Mohammad Ali Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," Computers in Biology and Medicine, vol.136, pp.104672, 2021.
[20] Mienye, Ibomoiye Domor, Yanxia Sun, and Zenghui Wang, "An improved ensemble learning approach for the prediction of heart disease risk," Informatics in Medicine Unlocked, vol.20, no.1, pp.100402, 2020.
[21] Dutta, Aniruddha, Tamal Batabyal, Meheli Basu, and Scott T. Acton, "An efficient convolutional neural network for coronary heart disease prediction," Expert Systems with Applications, vol.159, pp113408, 2020.
[22] Ishaq, Abid, Saima Sadiq, Muhammad Umer, Saleem Ullah, Seyedali Mirjalili, Vaibhav Rupapara, and Michele Nappi, "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," IEEE access, vol.9, no.1, pp. 39707-39716, 2021.
[23] 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.
[24] 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.
[25] 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.

BIOGRAPHIES OF AUTHORS

Monali Gulhane is recipient of young researcher award in 2021 by INSC, she has been merit holder in academics and received M.Tech in Computer Science Engineering from G. H. Raisoni College of Engineering and Technology for Women's (G.R.C.E.T.W), Nagpur, Maharashtra, in 2012-2014. She is currently pursuing PhD degree in Artificial Intelligence and Machine Learning from Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur.

T. Sajana received the B.Tech in Computer Science Engineering from Koneru Lakshmaiah College of Engineering (KLCE), Vaddeswaram,Guntur, in 2007. She received M.Tech (GATE) in Computer Science and Engineering from St.Ann's College of Engineering & Technology, Chirala, in 2009-11. She received PhD degree in Computer Science and Engineering from Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, in 2019. She is currently an Associate Professor in the Department of Computer Science and Engineering, KLEF, Vaddeswaram, Guntur. Her research interests include Machine learning, Data mining, Computational Intelligence, Deep learning.