Cardio meta-stack: a meta-classifier ensemble for enhanced cardiovascular disease prognosis

Swetha S.¹, Sneha Zolgikar¹, Manjula S. H.¹, Venugopal K. R.²

¹Department of Computer Science and Engineering, University of Visvesvaraya of Engineering (UVCE), Bangalore, India ²Former Vice Chancellor, Bangalore, India

Article Info

Article history:

Received May 13, 2024 Revised Aug 6, 2025 Accepted Nov 16, 2025

Keywords:

CVD DBSCAN Feature importance Risk factors Stacking meta-classifier

ABSTRACT

Cardiovascular diseases (CVDs) remain a significant global health concern, necessitating effective preventive measures and early diagnosis to reduce mortality rates. Leveraging machine learning models to identify risk factors holds great promise, especially in cardiology. This study introduces a robust methodology for prognosing cardiac illnesses based on patient-specific factors. By integrating five publicly available datasets from the UCI Repository and employing Feature Importance techniques for optimal risk factor selection, the proposed approach enhances prediction accuracy. Furthermore, the inclusion of the density-based spatial clustering of applications with noise (DBSCAN) algorithm assists in noise detection and removal, thereby improving model precision. The proposed Cardio Meta-Stack model, coupled with a stacking classifier ensemble, achieved an accuracy of 94.91%, surpassing that of traditional algorithms such as XGBoost 90.45%, demonstrating its efficacy in heart disease prediction.

This is an open access article under the CC BY-SA license.



1630

Corresponding Author:

Swetha S.

Department of Computer Science and Engineering, University of Visvesvaraya of Engineering (UVCE) Bangalore-560001, Karnataka, India

Email: swetha.s@uvce.ac.in

1. INTRODUCTION

Cardiovascular diseases (CVDs) present a significant global health challenge, ranking as the primary cause of both mortality and morbidity worldwide [1]. These encompass a spectrum of conditions affecting the heart and blood vessels, often resulting in profound consequences within healthcare systems [2]. They represent a crucial global health concern, accounting for a notable share of fatalities and disabilities across various regions. In both the United States and European countries, these diseases contribute significantly to mortality rates [3]. Statistical data reveals that CVDs are accountable for approximately one in every three deaths in the United States, underscoring their substantial impact on public health. Similarly, in European nations, CVDs frequently lead to high mortality rates, placing a significant burden on healthcare systems and society in general [4]. While earlier studies have explored the mortality and morbidity rates associated with CVDs, they have not explicitly addressed the influence of risk factor management on healthcare systems. Given their intricate interplay of risk factors, underlying mechanisms, and diverse outcomes, understanding CVDs holds paramount importance for healthcare systems and societies at large [5]. Numerous risk factors are associated with cardiovascular diseases, primarily categorized into modifiable and non-modifiable risk factors. Modifiable factors include behavioral aspects such as minimum physical activity, consumption of more sodium, excessive alcohol and tobacco consumption, metabolic conditions like raised blood pressure and high fasting plasma glucose levels, increased waist-to-hip-ratio ie; body-mass index (BMI), high levels of low-density lipoprotein (LDL), and diabetes and environmental factors like atmospheric air pollution [6].

ISSN: 2502-4752

These factors can be altered or managed to decrease the tendency of developing diseases. Non-modifiable risk factors encompass family medical history, gender, and age, which cannot be changed but can be managed through preventive measures.

In modern times, machine learning (ML) practices have revolutionized healthcare by providing powerful tools for data analysis, prediction and classification [7]. Within cardiology, ML has emerged as a critical tool for diagnosing, predicting, and managing cardiovascular diseases. The study [8] involves the use of two ML approaches, multi-layer perceptron (MLP) and K-nearest neighbor (KNN) to identify CVD. Experimental outcomes show that the MLP model outperforms, achieving an 82.47% detection accuracy and an area under the curve (AUC) value of 86.41% surpassing the performance of the KNN model. In a recent study, Noroozi *et al.* [9] examined the impact of multiple feature-selection strategies on cardiovascular-disease prediction and demonstrated that hybridized wrapper–filter combinations significantly improved model accuracy compared with single-method selection. Similarly, Xiong *et al.* [10] proposed a recursive feature-elimination and ensemble-learning hybrid for optimizing attribute relevance, reducing redundancy, and enhancing classifier performance. Fernandes and Miranda [11] investigated feature-importance ranking for early cardiovascular-risk assessment and highlighted that refined feature engineering strengthens diagnostic interpretability. These works collectively emphasize the importance of attribute-selection and feature-optimization techniques in developing efficient and reliable heart-disease prediction models.

In parallel, Shukla *et al.* [12] introduced a hybrid feature-selection and meta-ensemble framework combining classical algorithms for cardiovascular-risk assessment. Their approach demonstrated that optimized feature subsets enhance interpretability and reduce computational cost without relying on complex neural architectures. Likewise, Carnevale *et al.* [13] presented a feature-optimized random-forest-based model with parameter tuning for multimodal heart-disease prediction, achieving robust performance while maintaining model simplicity and transparency. These studies underline that carefully selected features and optimized traditional classifiers can provide strong, explainable, and computationally efficient alternatives for cardiovascular prognosis. In a study, by Abdullahi *et al.* [14] the focus was on utilizing advanced feature selection methods for predicting cardiovascular disease. The research compared the effectiveness of wrapper method and filter method with the wrapper method showing performance, in enhancing model accuracy.

The accuracy of a model depends on how well the dataset is preprocessed, incorporating various feature selection techniques and selecting the best model. Understanding CVDs is crucial for healthcare systems and societies due to their intricate interplay of risk factors, underlying mechanisms, and diverse outcomes. Existing methods of CVD prognosis often struggle with feature selection and model choice, leading to suboptimal predictive performance. This study addresses this gap by introducing cardio-meta stack, a metaclassifier ensemble approach that leverages stacking, feature importance, and outlier detection for enhanced feature selection. Our method demonstrates improved performance in CVD prognosis by: a) Selecting the most relevant features from a comprehensive dataset; b) Combining the strengths of multiple classifiers through stacking and meta-classifier selection; c) Identifying novel patterns and relationships in the dataset, demonstrating the effectiveness of cardio-meta stack in CVD prognosis.

The remaining section of paper is organized as follows. Section 2 covers detailed overview of our proposed cardio-meta stack model, including a problem statement, dataset description, and methodology. Section 3 delves into the results of our study, featuring an analysis of ML model stacking levels, their corresponding accuracies, and a comprehensive evaluation of our approach. Section 4 summarizes the key findings, highlighting the significance of our contributions, discussing the potential implications and future directions of our research.

2. PROPOSED MODEL

2.1. Problem statement

Despite significant advancements in medical science, cardiovascular diseases remain a prominent global threat to life. Identifying high-risk individuals in a timely manner and accurately prognosing cardiac illnesses are crucial tasks. However, prevalent risk assessment techniques often lack the necessary granularity to detect subtle yet impactful risk factors, leading to missed opportunities for early intervention.

2.2. Dataset description

The data set used is a fusion of 5 separate datasets: Cleveland, Hungary, Switzerland, VA Long Beach, and Statlog [15]. This merged dataset consists of 1190 patient records, encompassing 11 attributes and 1 target attribute. Within this dataset, a class value of 0 denotes the absence of heart disease, while a class value of 1 signifies its presence. To address missing values in the datasets, we implemented the mean replacement technique. Following this, we utilized the sklearn train-test split library to divide the dataset into training and test sets, ensuring an 80:20 split ratio. Table 1 gives complete description of analysis of dataset.

1632 □ ISSN: 2502-4752

Table 1. Exploratory analysis of dataset							
Parameter	Details						
Total count of data before train_test splitting	(Rows: 1190, Columns: 11)						
Selected features	Age, Sex, Chest pain type, resting bp, cholesterol, fasting blood sugar, max						
	heart rate, exercise angina, old peak and resting slope						
Excluded feature	Resting ECG						
Applying the density-based spatial	Identified and removed 6 outliers using DBSCAN on Age and max heart rate,						
clustering of applications with noise	FBS, and Resting BP						
(DBSCAN)							
Final processed dataset	The Finalized size (1184 records, 10 features)						

2.3. Cardio-meta-stack architecture-proposed architecture

Figure 1 illustrates the proposed stacking approach, depicting the suggested architecture. The study's implementation process includes various key steps aimed at efficiently predicting an individual's propensity for cardiovascular events based on relevant risk factors. Data importation and handling missing values through mean replacement technique. Following data preprocessing, the feature selection process is initiated to enhance model accuracy and reduce computational complexity. Feature selection is accomplished by employing the 'Feature Importance' technique, which identifies key features based on their attribute's p-score as shown in Figure 2. Each attribute is assigned feature importance scores, with higher scores indicating greater relevance to the final prediction. Outlier detection and elimination are carried out using the DBSCAN-based algorithm on the selected features. Classification of the dataset and prediction are performed using the XGBoost-based boosting algorithm [16]. Comparative analysis between the individual XGBoost classifier and the stacking technique is performed, selecting the classifier with the higher accuracy score.

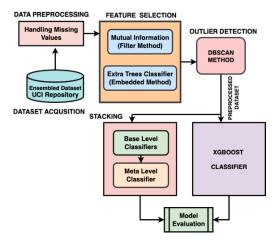


Figure 1. Proposed architecture

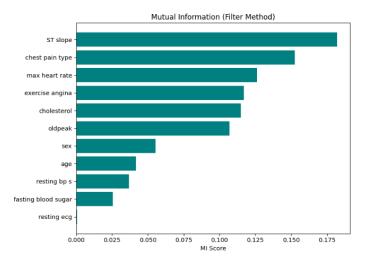


Figure 2. Feature importance scores

2.4. Five level stacking

Stacking, an ensemble learning method, constructs a novel model by merging the predictions from various nodes. This resultant model is then employed to make predictions on the test dataset. In the proposed study, a stacking approach is employed employing 5 classifiers as shown in Figure 3. This technique aids in alleviating biases and shortcomings inherent in individual models by consololidating their strengths, thereby augmenting predictive accuracy. The proposed stacking procedure cardio-meta-stack, is detailed in Algorithm 1.

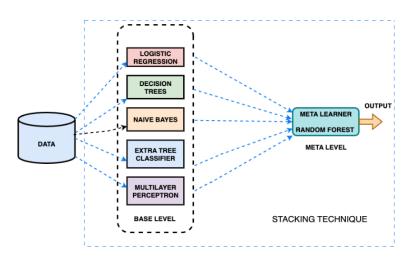


Figure 3. Level 5 stack model

Algorithm 1. Cardio-meta-stack

```
Input: Train (Dataset Training)
Input: Validation (Dataset Validation)
Input: Test (Dataset Testing)
Output: predictions (Predictions on Test set)
Models: S = \{RF, DT, NB, ET, NN\}
meta features = []
# Training base models
For each model type in S:
    Train base model model type on Train
    # Generating predictions on validation set
    preds val = Predictions of model type on Validation
    Append preds_val to meta_features
# Training meta-model
Train meta-model RF on meta features and Validation
# Generating predictions on test set
meta_features_test = []
For each model_type in S:
    preds test = Predictions of model type on Test
    Append preds test to meta features test
 predictions = Predictions of meta-model on meta features test
```

3. RESULTS AND DISCUSSION

3.1. Experimental setup

The key objective of this research is to efficiently forecast an individual's susceptibility to cardiovascular events based on relevant risk factors. All experiments were performed using Python 3.8 (Anaconda3) and OpenCV-Python.

3.2. Levels of ML model stacking and their accuracies

The stacking technique was employed with various levels and combinations of different models at each level. Random forest (RF) was chosen as the final estimator across all the levels. The highest accuracy

1634 □ ISSN: 2502-4752

of 94.91% was attained at level 5. The combinations of classifiers are as shown in Table 2. Despite exploring upto level 7, it is observed that level 5 provided the optimal balance between model complexity and predictive performance. Table 3 presents a comparative analysis of the proposed method with existing techniques. Table 4 summarizes the performance metrics, showing that the stacking model outperforms the XGBoost classifier. Figure 3 and Figure 4 presents the confusion matrix obtained by comparing 5 level stacking and XGboost algorithm. Figure 5 depicts the accuracy of both techniques.

We found that increasing the stacking level up to 5 correlates with improved accuracy, reaching 94.91% at level 5. The proposed method tends to have a higher proportion of accurate predictions as the stacking level increases, with level 5 providing the optimal balance between model complexity and predictive performance.

Table 2. Performance of ML model stacking at various levels

Levels of ML model stacking	ML models-> Final estimator RF	Accuracy (%)				
Level 1	LR	78.48				
Level 2	LR+DT	85.23				
Level 3	LR+DT+NB	86.92				
Level 4	LR+DT+NB+SVM	89.87				
Level 5	LR+DT+NB+ET+MLP	94.91				
Level 6	LR+DT+NB+ET+MLP+SVM	93.25				
Level 7	LR+DT+NB+ET+MLP+SVM+KNN	92.83				

Table 3. Comparision of proposed method with different techniques

Author and year	Technique	Acc. (%)	
Bashir et al. [17] 2021	Ensemble voting scheme	83	
Rani et al. [18] 2021	6 Hybrid classifiers	86.6	
Kavitha et al. [19] 2022	RF+DT+ Hybrid	88.7	
Saranya and Pravin [20] 2023	ya and Pravin [20] 2023 RF+FSFCA omitting 5 features		
	Without omitting features	86.1	
Alzubaidi et al. [21] 2022	Bagging ensemble	89.3	
Budholiya et al. [22] 2022	Hypertuning XGboost+Bayesian optimization	91.8	
Rajendran and Karthi [23] 2022	Ensemble (LR+NB)	92.7	
Ozcan and Peker [24] 2023	CART classification	87.2	
Uddin et al. [25] 2022	Boosting ensemble	91.1	
Proposed	Feature importance+5 level stacking technique	94.91	

Table 4. Performance metrics of stacking vs XGBoost

uoic	1. I chormance metrics of stacking vs 116B66								
	Metric (%)	Stacking (%)	XGBoost (%)						
	Accuracy	94.91	90.45						
	Precision	94.18	89.75						
	Recall	95.53	92.50						
	F1-score	94.97	91.13						

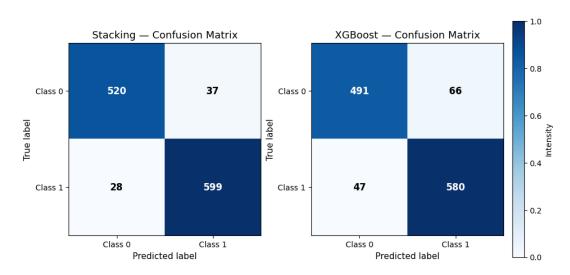


Figure 4. Confusion matrix of proposed stacking vs XGBoost

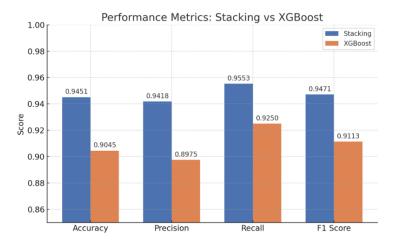


Figure 5. Comparision of XGBOOST and stacking models' accuracy results

4. CONCLUSION

In conclusion, our study successfully demonstrated the effectiveness of the RF based cardio metaclassifier approach in early-stage prediction of heart disease, with the ultimate goal of reducing mortality rates associated with this prevalent cause of death. By leveraging the feature importance approach and DBSCAN method, we identified significant risk factors and eliminated outliers, ensuring a robust dataset. Our ensemble of classifiers, trained on five publicly available datasets (Switzerland, Cleveland, Statlog, Hungarian, and Long Beach) from Kaggle, achieved superior performance over the XGB classifier, with accuracies of 90.45% and 94.91%, respectively. Our study suggests that using RF as the final estimator across all levels is associated with improved performance, unlike other techniques that may suffer from poor performance with increased complexity. The proposed method may benefit from further exploration of stacking levels and classifier combinations without adversely impacting accuracy. The implications of our findings are profound, offering a promising solution for early intervention and potentially averting fatalities resulting from delayed heart disease detection. Our approach can facilitate personalized risk assessment, enabling healthcare professionals to identify high-risk patients and provide timely interventions. Future research is to explore this approach to larger, more diverse populations and investigate the incorporation of additional risk factors.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	0	E	Vi	Su	P	Fu
Swetha S.	✓	✓	✓	✓	✓	✓	✓		✓	✓				
Sneha Zolgikar		\checkmark	✓			\checkmark	✓	\checkmark	\checkmark	\checkmark	✓			
Manjula S. H.	\checkmark			\checkmark						\checkmark	✓	\checkmark	\checkmark	
Venugopal K. R.					✓	✓				✓	✓	✓	✓	

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

1636 ☐ ISSN: 2502-4752

DATA AVAILABILITY

The data that support the findings of this study are openly available in UCI Machine learning Repository (Heart Disease Dataset).

REFERENCES

- [1] O. Gaidai, Y. Cao, and S. Loginov, "Global cardiovascular diseases death rate prediction," *Current Problems in Cardiology*, vol. 48, no. 5, p. 101622, May 2023, doi: 10.1016/j.cpcardiol.2023.101622.
- [2] S. Ghorashi *et al.* "Leveraging regression analysis to predict overlapping symptoms of cardiovascular diseases," *IEEE Access*, vol. 11, pp. 60254–60266, 2023, doi: 10.1109/ACCESS.2023.3286311.
- [3] K. K. Teo and T. Rafiq, "Cardiovascular risk factors and prevention: a perspective from developing countries," *Canadian Journal of Cardiology*, vol. 37, no. 5, pp. 733–743, May 2021, doi: 10.1016/j.cjca.2021.02.009.
- [4] C. W. Tsao *et al.* "Heart disease and stroke statistics—2023 update: a report from the american heart association," *Circulation*, vol. 147, no. 8, Feb. 2023, doi: 10.1161/CIR.00000000001123.
- [5] Consortium GCR, "Global effect of modifiable risk factors on cardiovascular disease and mortality," New England Journal of Medicine., vol. 389, no. 14, pp. 1273–85, 2023.
- [6] M. Swathy and K. Saruladha, "A comparative study of classification and prediction of cardio-vascular diseases (CVD) using machine learning and deep learning techniques," *ICT Express*, vol. 8, no. 1, pp. 109–116, Mar. 2022, doi: 10.1016/j.icte.2021.08.021.
- [7] A. Khan, M. Qureshi, M. Daniyal, and K. Tawiah, "A novel study on machine learning algorithm-based cardiovascular disease prediction," *Health & Social Care in the Community*, vol. 2023, pp. 1–10, Feb. 2023, doi: 10.1155/2023/1406060.
- [8] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Medicine*, vol. 17, no. 1, pp. 1100–1113, Jun. 2022, doi: 10.1515/med-2022-0508.
- [9] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Scientific Reports*, vol. 13, no. 1, p. 22588, Dec. 2023, doi: 10.1038/s41598-023-49962-w.
- [10] X. Xiong, S. Zhong, and Y. Lu, "Separable and reversible data hiding scheme for medical images using modified Logistic and interpolation," *Biomedical Signal Processing and Control*, vol. 87, p. 105521, Jan. 2024, doi: 10.1016/j.bspc.2023.105521.
- [11] R. S. Fernandes and J. G. V. Miranda, "An agent-based model for studying the temperature changes on environments exposed to magnetic fluid hyperthermia," *Computers in Biology and Medicine*, vol. 170, p. 108053, Mar. 2024, doi: 10.1016/j.compbiomed.2024.108053.
- [12] P. K. Shukla, N. N. Singh, and M. R. Kumar, "Hybrid feature selection and meta-ensemble framework for cardiovascular risk assessment using classical machine learning algorithms," *IEEE Access*, vol. 12, pp. 290–302, doi: 10.1109/ACCESS.2024.3508902.
- [13] L. Carnevale et al. "Towards a precision medicine Solution for optimal pediatric laparoscopy: an exploratory data analysis for features Selections," Biomedical Signal Processing and Control, vol. 88, p. 105321, Feb. 2024, doi: 10.1016/j.bspc.2023.105321.
- [14] A. Abdullahi, M. Ali Barre, and A. H. Elmi, "A machine learning approach to cardiovascular disease prediction with advanced feature selection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, p. 1030, 2024.
- [15] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease Dataset," *Heart Disease*, 1989. https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
- [16] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine learning technology-based heart disease detection models," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, Feb. 2022, doi: 10.1155/2022/7351061.
- [17] S. Bashir, A. A. Almazroi, S. Ashfaq, A. A. Almazroi, and F. H. Khan, "A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction," *IEEE Access*, vol. 9, pp. 130805–130822, 2021, doi: 10.1109/ACCESS.2021.3110604.
- [18] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, Sep. 2021, doi: 10.1007/s40860-021-00133-6.
- [19] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in 2021 6th International Conference on Inventive Computation Technologies (ICICT), Jan. 2021, pp. 1329–1333, doi: 10.1109/ICICT50816.2021.9358597.
- [20] G. Saranya and A. Pravin, "A novel feature selection approach with integrated feature sensitivity and feature correlation for improved prediction of heart disease," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 9, pp. 12005–12019, 2023, doi: 10.1007/s12652-022-03750-y.
- [21] M. Alzubaidi, M. Al-Amidie, O. Al-Asadi, and J. S. ia, "An ensemble-based approach for heart disease prediction using bagging and decision trees," *Applied Computing and Informatics*, vol. 18, no. 3, pp. 321–332, 2022.
- [22] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013.
- [23] R. Rajendran and A. Karthi, "Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers," *Expert Systems with Applications*, vol. 207, p. 117882, Nov. 2022, doi: 10.1016/j.eswa.2022.117882.
- [24] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," Healthcare Analytics, vol. 3, p. 100130, Nov. 2023, doi: 10.1016/j.health.2022.100130.
- [25] S. Uddin, M. Khan, and A. Hossain, "Boosting-based predictive modeling for cardiovascular disease diagnosis," BMC Medical Informatics and Decision Making, vol. 22, no. 101, pp. 1–12, 2022.

BIOGRAPHIES OF AUTHORS



Swetha S. is a research scholar from the Department of Computer Science and Engineering, University of Visvesvaraya College of Engineering, Bengaluru, India. She completed her master's in Engineering in Computer Science and Engineering at the University of Visvesvaraya College of Engineering, Bengaluru, India. Completed her bachelor's in the same branch from Government Sri Krishna Rajendra Silver Jubilee Technological Institute, Bengaluru, India. Currently, she is a full-time research scholar pursuing the specialization "Optimized Machine Learning and Deep Learning Techniques for Prediction of Heart Disease". Her subject interests include AI, machine learning, and deep learning applications. She can be contacted at email: swetha.s@uvce.ac.in.



Sneha Zolgikar () so has a bachelor of Engineering from Visvesvaraya Technological University and a Master of Technology in Computer Networking from University Visvesvaraya College of Engineering. Her areas of interest in research are prompt engineering, AI, and machine learning. She can be contacted at email: zolgikar25@gmail.com.



