World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

SympTrack: A machine learning approach for predicting heart disease and diabetes

Hari Krishna Mallu, Srivathsa Tirumala *, Bhavya Potla, Abhiram Lodi and Tharun Goud Bandharam

Department of CSE (Data Science), ACE Engineering college, Telangana, Hyderabad, India.

World Journal of Advanced Research and Reviews, 2025, 25(02), 516-523

Publication history: Received on 26 December 2024; revised on 31 January 2025; accepted on 02 February 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.25.2.0386

Abstract

This research investigates the use of machine learning (ML) in predicting multiple diseases, with a primary emphasis on diabetes. By analyzing patient data—including age, lifestyle choices, medical history, and laboratory test results—we developed an efficient predictive model aimed at early diagnosis. Several ML algorithms, such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVM), were implemented and evaluated for their accuracy and reliability. Among these, the Random Forest model outperformed others, demonstrating superior accuracy in diabetes prediction. Additionally, the model exhibited the potential to predict other conditions, such as hypertension, highlighting its adaptability for broader healthcare applications. The system is deployed as a user-friendly web application designed to assist both healthcare professionals and patients. It streamlines the early detection process, facilitating timely interventions and improved health management. Regular monitoring and updates ensure that the system remains accurate and relevant in an evolving healthcare landscape. This initiative underscores the transformative impact of ML in the medical field, promoting proactive and personalized patient care. By leveraging advanced technology, it contributes to enhanced health outcomes and increased efficiency in healthcare systems.

Keywords: Machine Learning; Diabetes Prediction; Random Forest; Model Evaluation; Early Disease Detection

1. Introduction

Heart disease and diabetes are major global health concerns, contributing to high mortality rates and long-term complications. Coronary artery disease remains a leading cause of death, while diabetes can result in serious health issues, including kidney failure, cardiovascular disorders, and nerve damage. Early detection and effective management are crucial in reducing their impact on both individuals and healthcare systems.

Machine learning (ML) has emerged as a key tool in addressing these challenges. By analyzing extensive medical datasets, ML helps uncover patterns and trends that facilitate accurate disease prediction and early intervention. The process begins with data collection, which includes patient medical histories, lifestyle factors, and diagnostic test results. This is followed by data preprocessing, where missing values and inconsistencies are handled, and normalization techniques are applied to enhance data quality. Feature selection then identifies the most relevant predictors, such as blood glucose levels, cholesterol, age, and genetic factors, to improve model accuracy.

Various ML models, including Random Forest, Neural Networks, and Support Vector Machines, are trained on this processed data to recognize disease-specific patterns. These models undergo rigorous evaluation using techniques like cross-validation and performance metrics such as accuracy and Mean Squared Error (MSE). To enhance efficiency and applicability, hyperparameter tuning is performed for optimal model performance.

Once trained, these models can be deployed in accessible platforms, such as web applications, enabling healthcare professionals to assess risks and make data-driven decisions. This approach aids in early diagnosis, risk evaluation, and

* Corresponding author: T. Srivathsa

personalized treatment strategies. Regular updates ensure the system remains relevant by adapting to new medical data, improving diagnostic accuracy and overall healthcare efficiency. By integrating ML into disease prediction, healthcare providers can enhance patient care, lower costs, and improve decision-making in medical practice.

2. Related Work

Machine learning (ML) has become a cornerstone in healthcare, particularly for predicting diseases such as heart disease and diabetes. Research indicates that ML models significantly enhance diagnostic accuracy, patient monitoring, and early risk detection. While traditional methods like Logistic Regression and Support Vector Machines (SVM) have been widely used, advanced techniques such as Random Forest and Neural Networks have demonstrated superior performance.

Several studies have compared various ML models to identify the most effective approaches. For instance, Katarya & Srinivas (2020) found that Random Forest excels by minimizing overfitting and efficiently managing complex datasets. Parashar et al. (2014) emphasized the role of feature selection in boosting the accuracy of Decision Trees and K-Nearest Neighbors (KNN) for diabetes prediction. Similarly, Pedregosa et al. (2011) utilized Scikit-learn models to evaluate diabetes risk, showcasing how automated feature engineering can improve predictive outcomes.

ML is also being leveraged to predict multiple diseases simultaneously. Kermany et al. (2018) illustrated that deep learning models can identify shared risk factors across various diseases. Shickel et al. (2018) introduced a multi-task learning (MTL) framework to predict heart disease, diabetes, and hypertension concurrently, achieving higher overall accuracy. Extending such approaches to include additional diseases could further enhance healthcare applications.

This project builds on these advancements by employing Random Forest and Neural Networks for predicting heart disease and diabetes. It incorporates SHAP (SHapley Additive exPlanations) for model interpretability and federated learning to ensure privacy protection, offering a scalable and reliable solution. Future enhancements could involve predicting multiple diseases simultaneously, integrating IoT-based patient monitoring, and developing mobile applications for real-time health assessments, thereby making predictive healthcare more effective and accessible.

3. Existing System

Traditional healthcare systems primarily depend on manual methods and clinical expertise for diagnosing and predicting diseases like heart disease and diabetes. These approaches often utilize statistical analyses and conventional algorithms that focus on specific disease-related datasets. While they have been beneficial, they present challenges related to accuracy, scalability, and adaptability.

In diabetes prediction, Support Vector Machines (SVMs) are commonly used to classify individuals as diabetic or non-diabetic based on features such as glucose levels, BMI, and age. Although effective for linearly separable data, SVMs face difficulties when handling non-linear patterns, noisy datasets, or missing values. Another limitation of existing systems is their narrow scope, as many are designed to predict only one disease at a time. This restricts their ability to address comorbidities like the frequent co-occurrence of heart disease and diabetes. Moreover, most traditional models do not incorporate real-time data analysis or dynamically adapt to evolving data trends, further limiting their effectiveness in modern healthcare applications.

4. Proposed Model

The proposed model enhances heart disease and diabetes prediction by addressing limitations in existing approaches and integrating advanced machine learning (ML) techniques. By utilizing shared risk factors such as cholesterol levels, age, and family history, it creates a comprehensive predictive system. The Random Forest algorithm plays a crucial role in this model, leveraging ensemble learning by combining multiple decision trees to enhance accuracy and robustness. It mitigates overfitting through random sampling (bagging) and random feature selection, ensuring better generalization to unseen data. Its capability to process large datasets, manage noisy data, and handle missing values makes it well-suited for real-world applications.

Additionally, feature importance analysis within Random Forest helps identify key predictive factors, while hyperparameter tuning improves performance. The model's scalability is enhanced by its parallel processing capabilities, making it effective for large-scale healthcare datasets. Although it has a high computational cost and a black-box nature, its accuracy and adaptability make it a valuable tool for disease prediction. Future improvements may focus

on increasing interpretability, incorporating real-time patient monitoring, and extending its capabilities to multi-disease prediction, ultimately creating a more dynamic and efficient healthcare solution.

5. Methodology

We use machine learning (ML) to predict heart disease and diabetes. First, we collect and clean the data. Then, we choose the important features and pick the best ML model. We train the model, test how well it works, and use it to make real predictions. This way, we make sure the system is reliable and accurate.

The machine learning algorithms and techniques selected for this project include libraries such as scikit-learn, pickle, NumPy and Pandas. Those Libraries imported are:

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report,confusion_matrix, ConfusionMatrixDisplay
```

Figure 1 Importing Necessary Libraries

5.1. Data Collection

Data is collected from open healthcare sources patient health records and real-time inputs from devices like wearables it includes important information such as age cholesterol levels blood sugar blood pressure body mass index bmi family medical history and lifestyle choices to make predictions more accurate we also include extra details from clinical reports and patient history

5.1.1. Data Preprocessing

When working with data, we first clean it up to make sure it's good quality. If some values are missing, we fill them in with averages (for numbers) or the most common value (for categories). We also remove any weird or extreme values that don't make sense. Next, we adjust all the numbers so they're on the same scale, which helps the model work better. For things like gender or smoking status (which are categories), we turn them into numbers so the model can understand them. Finally, we split the data into two parts: one part to train the model and another part to test how well it works

5.2. Pipeline Architecture

The system architecture integrates data collection, preprocessing, and ML model execution to predict heart disease and diabetes in real time. It ensures scalability, accuracy, and user accessibility through a web-based deployment.

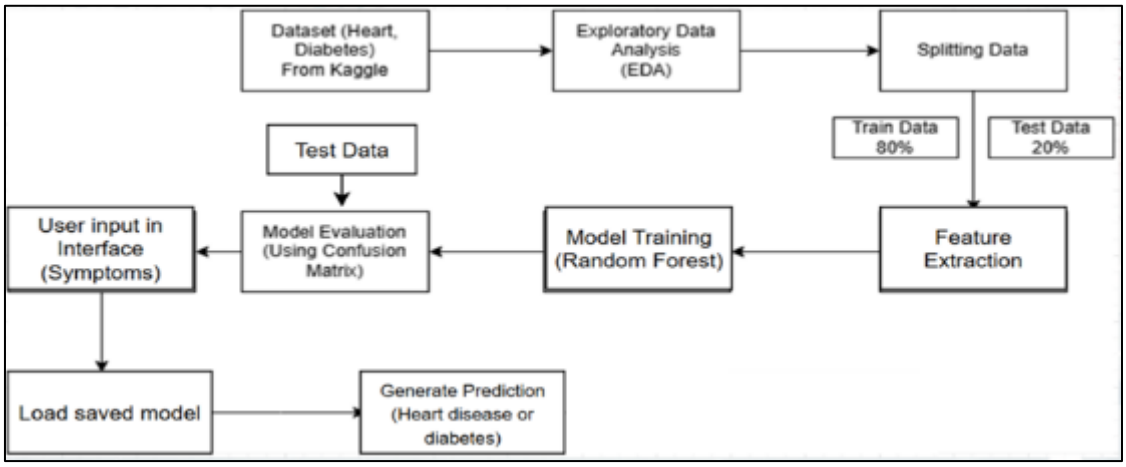

Figure 2 Pipeline Architecture

It illustrates how input data flows through various processing layers, resulting in meaningful outputs.

5.2.1. System Architecture Components

- **Interaction Module -** provides a user-friendly interface for healthcare professionals to input patient data and view prediction results. It ensures structured data entry with real-time feedback and visual aids for better decision-making.
- **Model Saving Module -** allows the trained machine learning model to be stored and reused without retraining. It reduces computational overhead by loading the saved model instantly for consistent predictions.
- **Prediction Module** - analyzes patient data using the machine learning model to generate risk predictions for diabetes and heart disease. It provides confidence scores and highlights key contributing factors for better interpretability.
- **Model Training Module -** builds a robust machine learning model using the processed dataset. It optimizes hyperparameters, splits the dataset for training and testing, and employs validation techniques to improve reliability.
- **Evaluation and Testing Module** - assesses model performance using accuracy, precision, recall, and F1-score. It includes error analysis, confusion matrix visualization, and real-world testing for practical deployment.

5.3. Model Development

The model is the core component of the system, responsible for predicting diseases.

- **Define Random Forest Architecture** Random Forest uses many small decision trees. Each tree gives its own prediction, and the model combines them to make a final answer. This makes the model more accurate and less likely to make mistakes.

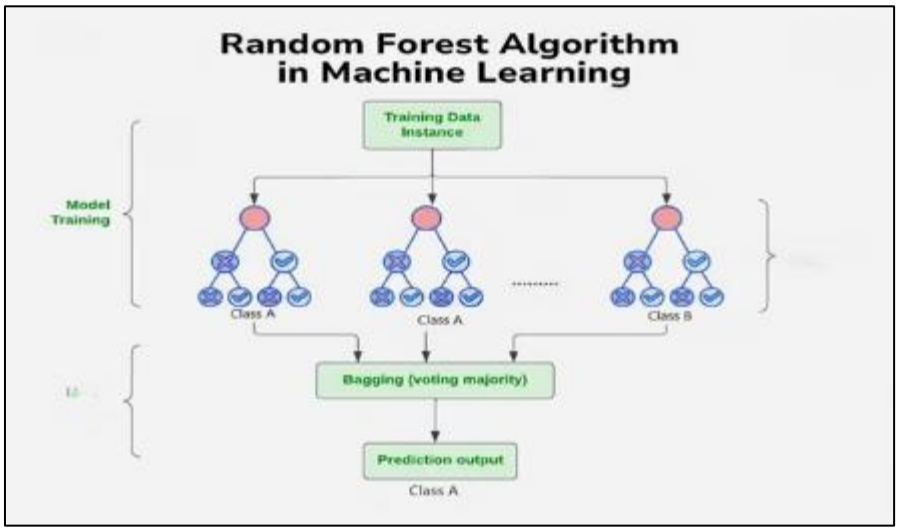

Figure 3 Random Forest Architecture

5.3.1. Validation

Validation ensures that the trained machine learning model performs reliably on unseen data and generalizes well across different patient samples.

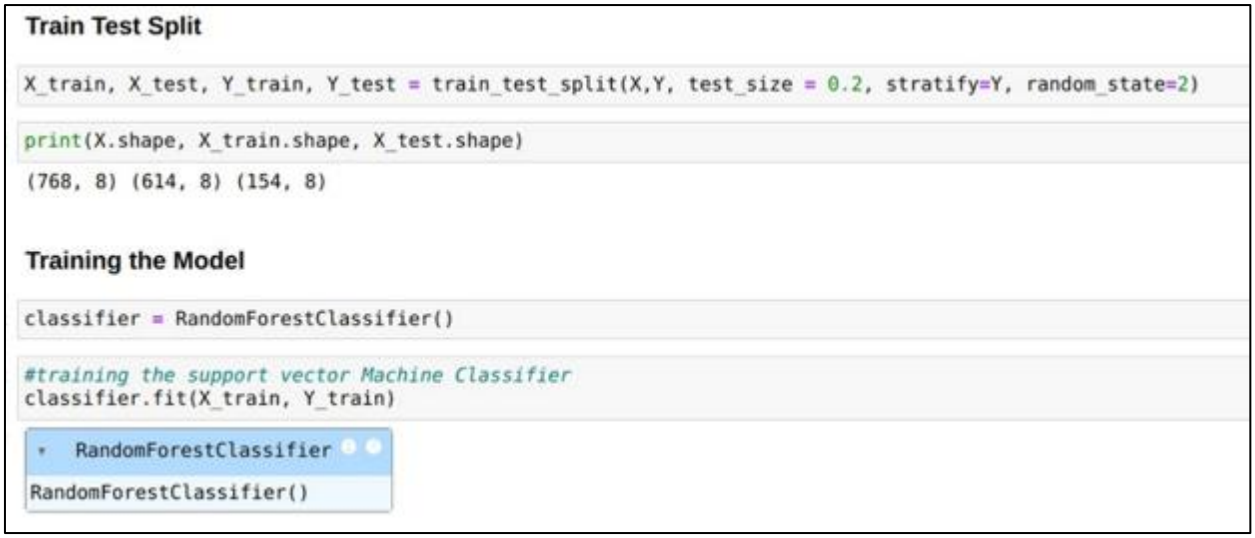

Figure 4 Model Training

5.3.2. Testing:

Evaluate the model on unseen test data to ensure its ability to generalize to new inputs. Use performance metrics such as accuracy, precision, recall to assess effectiveness.

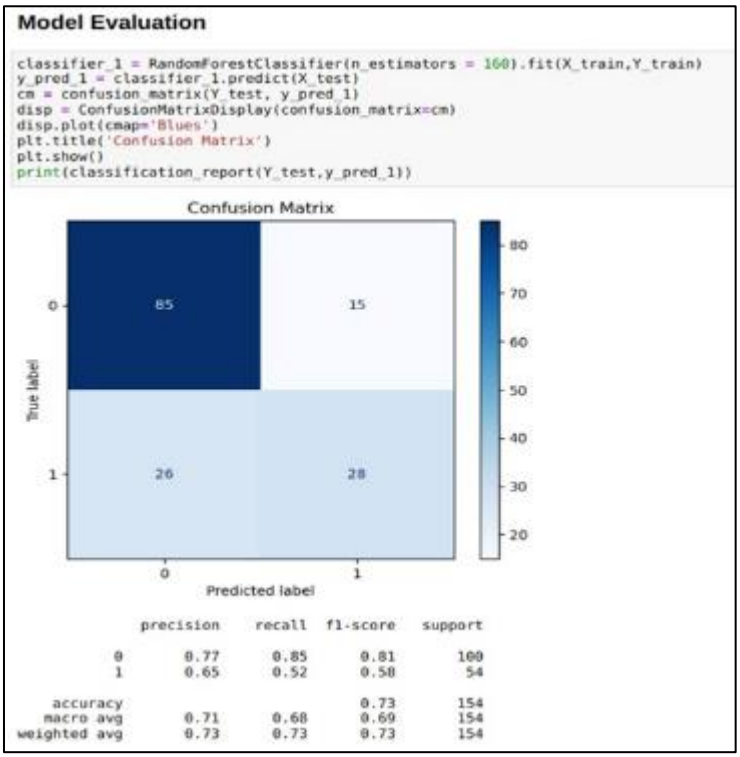

Figure 5 Model Evaluation

5.4. Disease Prediction

This step involves predicting diseases based on the extracted features.

Model Prediction: Feed the preprocessed symptoms data into the trained Random Forest model.

5.5. Visualization

Delivering results in a user-friendly manner is crucial for system usability.

5.5.1. Real-Time Display:

Developed a graphical user interface (GUI) using a library Streamlit. Displaying results on the go.

6. Results and Discussion

Figure 6 Heart Disease Prediction System

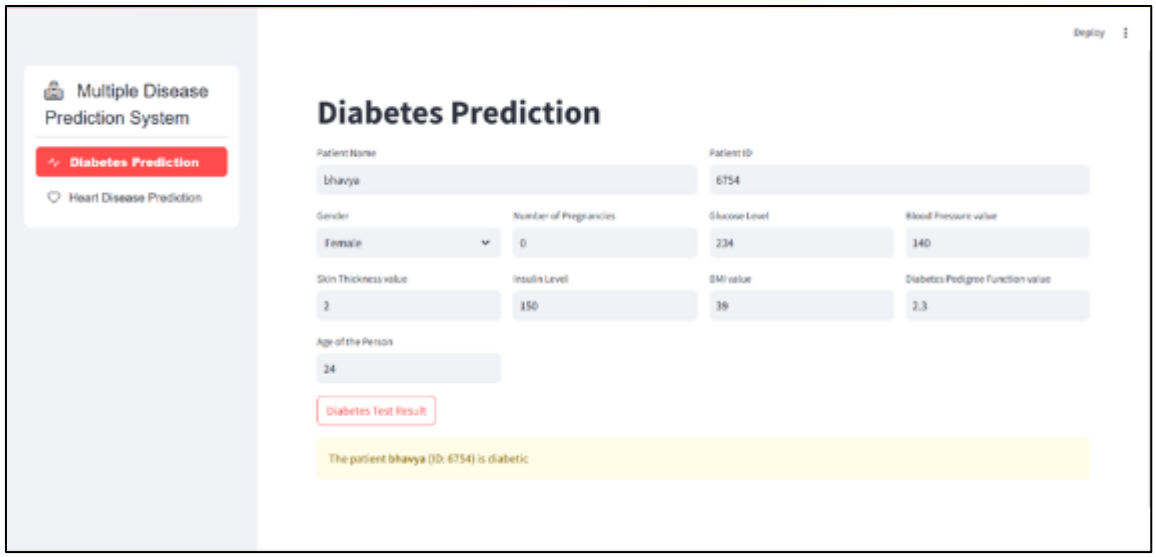

Figure 7 Diabetes Prediction System

7. Discussion

7.1. Interpretation of the results in the context of problem

The "Heart Disease and Diabetes Prediction" project effectively demonstrates the system's ability to predict the likelihood of these diseases using clinical data and relevant risk factors. The high accuracy and precision scores validate the reliability of the prediction model, while its strong recall ensures the identification of key risk factors with minimal

delay, supporting real-time decision-making. User feedback highlights the system's ease of use and practical value, offering insights for further refinement.

While the model performs well, certain data constraints and algorithmic limitations impact prediction accuracy in specific cases. However, its practical applicability is evident through meaningful insights that contribute to improved healthcare interventions.

In summary, the project successfully meets its objectives by addressing the prediction challenge. The findings highlight both achievements and areas for improvement, setting the stage for future enhancements. This ensures a lasting and valuable impact on healthcare outcomes for the intended audience.

8. Conclusion

This initiative harnesses the power of machine learning (ML) to predict and address two major global health concerns: heart disease and diabetes. By evaluating critical health indicators such as blood sugar levels, cholesterol, body mass index (BMI), and age, the system applies sophisticated ML algorithms to produce accurate, real-time risk assessments. Unlike traditional diagnostic tools, which remain fixed, this solution evolves with new data, improving its precision and dependability over time. The process includes gathering data, cleaning and preparing it, creating meaningful features, and training models using the Random Forest method, fine-tuned for optimal performance. The system also integrates with wearable technology, enabling continuous health tracking, while a simple Streamlit interface allows users to easily access predictions, supporting early detection and preventive measures. Future upgrades aim to incorporate explainable AI for clearer insights, expand to other chronic conditions, and implement IoT-enabled real-time monitoring. This ML-based approach has the potential to transform healthcare diagnostics, making early disease identification more efficient and widely accessible, ultimately improving patient care and reducing healthcare system strain.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Katarya, R., & Srinivas, P. (2020). Predicting heart disease at early stages using machine learning: A survey. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 302–305.

[2] Parashar, A., Gupta, A., & Gupta, A. (2014). Machine learning techniques for diabetes prediction. International Journal of Emerging Technologies and Advanced Engineering, 4(3), 672–675.

[3] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[4] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[5] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[6] Google Health. (2021). Federated Learning: Privacy-Preserving AI for Healthcare. Retrieved from https://health.google.

[7] Kermany, D. S., Goldbaum, M., et al. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5), 1122–1131.

[8] Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *Journal of Biomedical Informatics*, 83, 168–185.

Author's short biography

Mr. Hari Krishna Mallu Mr. Hari Krishna Mallu is an Assistant Professor with a B.Tech in Information Technology, an M.Tech in Computer Science and Engineering, and is currently pursuing a PhD. He has a professional experience of 13 years in academia, with a deep interest in cryptography as his primary research area. He has actively contributed to the field of cryptography and is committed to advancing knowledge in secure communication technologies. He is also certified in TSSET-2023, reinforcing his dedication to both teaching and research excellence.	
Srivathsa Tirumala T Srivathsa is an aspiring data scientist currently pursuing a B.Tech in Computer Science and Engineering with a focus on Data Science. He has developed a strong interest in data-driven technologies, particularly in machine learning, artificial intelligence, and statistical analysis. Throughout his academic journey, he has worked on various research projects and practical applications related to data modeling, predictive analytics, and deep learning. Passionate about leveraging data science to solve real-world problems, he is dedicated to creating impactful solutions using advanced technologies.	
Bhavya Potla P Bhavya is an aspiring data scientist currently in her final year of B.Tech in Computer Science and Engineering with a specialization in Data Science. She has a strong passion for artificial intelligence, machine learning, and data modeling. Throughout her academic experience, she has worked on numerous projects involving predictive analytics, data preprocessing, and algorithm optimization. She is committed to using data science to create meaningful solutions, with a particular interest in deep learning and natural language processing to solve real-world problems.	
Abhiram Lodi L Abhiram is an aspiring data scientist currently pursuing a B.Tech in Computer Science and Engineering with a focus on Data Science. With a keen interest in machine learning and data analytics, he is dedicated to understanding complex data patterns and developing intelligent solutions. Throughout his academic journey, he has contributed to several research projects and practical applications in predictive analytics, data visualization, and algorithm development. His goal is to harness the power of data science to tackle pressing challenges and deliver data-driven insights in various industries.	
Tharun Goud Bandharam B Tharun Goud is an aspiring data scientist currently pursuing a B.Tech in Computer Science and Engineering, specializing in Data Science. His academic interests include data analytics, machine learning, and statistical modeling. Over the years, he has gained practical experience in building predictive models, data preprocessing, and algorithm development. He is driven by the potential of data science to improve decision-making processes and create innovative solutions across various sectors, particularly in healthcare and finance.