CARDIOMIND: AI-DRIVEN HEART DISEASE PREDICTION FOR BETTER HEALTH OUTCOMES

Divya D[a]*, MayaEapen[b] , Vanitha R[c] and Jetlin C.P[d]

[a]Associate Professor, Department of Computer Science and Engineering, Jerusalem College of Engineering,divya21cs@gmail.com

[b]Professor, Department of Computer Science and Engineering, Jerusalem College of Engineering, mayaeapen@jerusalemengg.ac.in

[c]Associate Professor, Department of Cyber Security, Jerusalem College of Engineering, vanithasenthilemail@gmail.com

[d]Assistant Professor, Department of Computer Science and Engineering, Jerusalem College of Engineering, jetlin@jerusalemengg.ac.in

Abstract--Heart disease continues to be a major problem for public health all over the world, which calls for techniques of prediction that are both accurate and fast. Within the scope of this research project classify the data in binary form, predictive analytics techniques are used to find presence or absence of cardiac disease. In this work, we make use of ensemble learning method, maximum-margin classifier, and log-linear classifier. Numerous datasets are used using these techniques that include features such as age, gender, the kind of chest pain, and physiological indicators such as blood pressure and cholesterol levels. The project investigates the difficulties associated with unbalanced datasets, data privacy, and the interpretability of models, all while working towards improving the accuracy of data predictions. When dealing with the processing of sensitive health data, ethical issues and regulatory compliance are of the utmost importance. The results provide a contribution to the ever-changing world of predictive healthcare analytics, highlighting the need to strike a balance between accuracy, privacy, and ethical issues in models that forecast cardiovascular disease.

Keywords: *heart disease prediction, machine learning, binary classification, imbalanced datasets, data privacy, model interpretability, cardiovascular disease.*

INTRODUCTION

Given that cardiac illness remains such biggest contributors of morbidity and mortality worldwide, advanced techniques for early detection and prediction by healthcare providers are necessary. In this project, we investigate how predictive analytics techniques may improve the accuracy of coronary artery bypass graft prediction.

As we engage in this research, we recognize the problems provided by unbalanced datasets, highlighting the need for correcting biases to maintain the robustness of our prediction models. Moreover, issues of data privacy and ethical treatment of sensitive health information are crucial, necessitating strict measures to conform to legal requirements, such as those mandated by HIPAA. Model interpretability emerges as a focus element, acknowledging the necessity of clear and intelligible forecasts in essential healthcare decision-making.

This work contributes to the ongoing discussion on the moral implications of using, predictive analytics techniques in healthcare in addition to attempting to boost the fidelity of cardiac illness prototypes. Through navigating such complex interplay by predictive ability, data security, and moral considerations, we want to provide important perspectives that might responsibly and meaningfully impact the creation and use of heart disease models for forecasting.

I. LITERATURE SURVEY

A. Related Works

Abishek et al. studied the numerous difficulties have been conducted that highlight the difficulties in predicting cardiac disease and investigate cutting-edge strategies like genetic algorithms and backpropagation. Comparative evaluations demonstrate the advantage of K-Nearest Neighbours. Efforts also include effective data extraction for prediction system optimisation. The body of research supports KNN's choice for precise and effective cardiac disease prediction by highlighting its flexibility, speed, and dependability.[1]

In order to predict heart attack and stroke, Farzana et al. conducted a global study on the impact of cardiovascular diseases using predictive analytics techniques . Many algorithms, including maximum-margin classifier, ensemble learning method, are studied. predictor selection boosts the performance of the paradigm; Random Forest and PCA attain the highest accuracy of 92.85%, indicating noteworthy progress in early diagnosis and intelligent healthcare systems.[2]

Simran verma et al. implemented the growing worldwide problem of cardiovascular diseases (CVDs) highlights the role that prediction techniques play in lowering death rates. It promotes using machine learning and data mining methods to glean insightful information from healthcare databases. The research emphasises how important it is to forecast heart disease accurately and recommends investigating other machine learning techniques for improved analysis and early illness detection.[3]

According to a study by Vijeta et al., heart-related illnesses are become more prevalent, which emphasizes the need of early identification. Using a UCI benchmark dataset, the study employs machine learning (ML) methodologies to predict cardiac illness using techniques including ensemble learning method, maximum-margin classifier. The data shows that ensemble learning outperforms other predictive analytics techniques with an accuracy rate of 99%. The research demonstrates how predictive analytics might act as a decision support system for physicians.[4]

The intricacy of diagnosing cardiac illness as well as the use of medical judgment assistance systems (DSS) that include data mining methods were examined by Kelibone et al. It raises concerns about the dependence on overall accuracy when evaluating classifiers and emphasizes the need of taking class-level accuracy into account. The comparison analysis highlights the need of giving preference to models that exhibit more accuracy across the board, including at the class level.[5]

In order to foresee the probability of cardiovascular disease, Santhana Krishnan.J applies information mining methodologies through the implementation of choice tree and bayes's algorithms. The choice tree model exhibits a higher level of accuracy, 91%, in comparison to the Bayes model (87%). In addition to recommending choice tree categorization for medical datasets, the study investigates additional predictive algorithms and potential extensions for outsourcing the analysis of cardiac conditions. [6]

Mohammed Jawwad Ali Junaid developed an early-stage heart disease prediction, emphasising the ongoing nature of the condition's progression and the importance of lifestyle choices. It presents a hybrid model that predicts and suggests preventive actions for cardiac patients using data science techniques (Naïve Bayes, ANN, and SVM). The hybrid model has improved sensitivity (91.47%), specificity (82.11%), and accuracy (2%). The study looks at things like strain, physical exercise, and heredity. The hybrid algorithm's effectiveness is emphasised in the conclusion, which also suggests applications for data science and smart devices to improve heart illness detection and awareness.[7]

The increasing prevalence of cardiac diseases in the modern era was investigated by Mamatha et al., who also highlighted the difficulties of predicting these conditions in the absence of medical diagnostics. The study employed attributes for maximum-margin classifier, ensemble learning method, algorithms for categorization in order to diagnose cardiac diseases by utilizing data mining techniques. Utilizing data obtained from Jubilee Mission Hospital, the results underscore the effectiveness of ANN in achieving a high degree of precision in the diagnosis of cardiac disease. Their objective is to offer cost-effective insights for early detection and preventive measures. [8]

II. METHODOLOGY

The method includes gathering datasets, preprocessing them, training models using SVM, Random Forest, and Logistic Regression, and creating Flask applications for forecasting heart disease.

1. Ensemble learning method

The ensemble learning method is a subset of the dataset is used to train each tree, and the individual forecasts made by each tree are combined to produce the final prediction. The durability of the model is increased, and overfitting is prevented, by using this ensemble approach. Since Random Forest can handle complex datasets and catch subtle patterns within the feature space, it is particularly suitable for our assignment on heart disease prediction. The algorithm is effective in healthcare applications because of its ability to prioritise different qualities, which enhances interpretability.

2. Maximum-margin classifier

A versatile technique that can cope with both linear as well as non-linear judgment limits is the maximum-margin classifier. Finding the most suitable feature space that maximizes the margin between different classes aim of maximum-margin classifier. This margin maximizing enhances the system's generalization ability and makes the approach more appropriate for binary classification issues such as heart disease prediction. MMC is capable of handling nonlinear connections and high-dimensional data, which allows it to provide accurate and dependable models for our predictive analytics.

3. Log-linear classifier

A classification method called a log-linear classifier is used to forecast the likelihood that an instance will belong to a certain class. It simulates the association between one or more independent factors (health characteristics) and a binary dependent variable (heart disease presence or absence). The logistic function is used by the log-linear classifier to estimate the probability, making it easier to comprehend the findings. The log-linear classifier offers a elementary and effortlessly comprehensible strategy for determining the probability of heart illness based on input data in the context of cardiac illness. It is a useful member of the group of algorithms used in the project because of its simplicity and effectiveness. The accuracy and dependability of the cardiac ailment identification system are enhanced by the combined use of these techniques. Each method is chosen based on how well it performs in managing distinct elements of the dataset and patterns related to cardiac disorders. For deployment in the Flask-based web application, a thorough and efficient predictive model is guaranteed by the ensemble technique and meticulous assessment.

A. FEATURE ATTRIBUTE

Every characteristic offers crucial information about the health characteristics pertinent to our prediction quest. The dataset for heart disease prediction contains a wide range of health parameters. Age, gender, blood pressure, cholesterol, kind of chest pain, and ECG results may all have an impact on heart health. Exercise-induced angina, thallium discoveries, and fasting blood sugar are examples of binary indicators. Peak heartbeat, the sadness, ventricular gradient, and vascular fluoroscopy data are examples of quantitative measures. The target variable indicates the presen

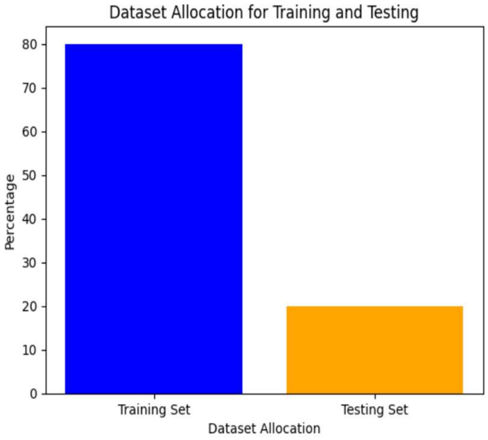

Fig. 1 Dataset Allocation

Then divided the raw data into 80% to be trained, and 20% to be examined throughout the modelling and assessment phases. This allowed us to efficiently gauge the performance of our model.

Table I: Attributes

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It allows visualization of the performance of an algorithm.

Characteristic Component	**Summary**
Years of old	A person's ages
Genre	Male or female identity
Cause of Cardiac Illness	Type of chest pain experienced (Categorical)
BP	measuring arterial bp
Fat	Measures of fat
Plasma Levels During a Diet	Higher than 120 mg/dl of glucose in the blood when fasting (Binary)
EKG Abnormalities	Electrocardiogram results (Categorical)
Highest cardiac Beat	Reached peak pulse of heartbeat
Angina Caused by Workout	The appearance of coronary provoked by exercising (Binary code)
ST Gloom	Exercise-induced depressive disorders in comparison with relaxation
ST Segmentation Gradient	Gradient within the cardiac wave's ST region.
Amount of Fluro Tubes	Primary artery count as determined by fluoroscopy colour
The thallium	The element thal endurance test outcomes (categorical)
Cardiac Conditions	Whether cardiac dysfunction (Binary code) is present or not

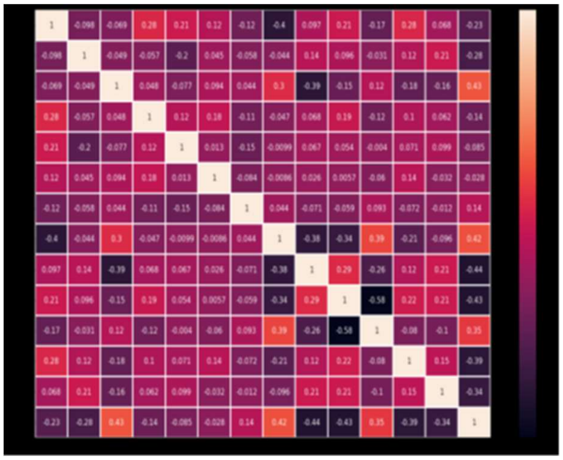

Fig:2 Confusion Matrix

III. PROPOSED SYSTEM DESIGN

In order to determine the thorough comparison of the method and create a system that uses predictive analytics to determine an individual's likelihood of cardiovascular disease, we are using three distinct sophisticated methods of predictive analytics in this project. Resources and functionality for building websites using the python employing an inexpensive browser environment. User-inputted health data is processed by a trained predictive analytics algorithm, which outputs predictions. Preprocessing, safe data management, a user-friendly interface, and reliable data are all necessary to guarantee correctness and privacy. It encourages users to make better health-related decisions and provides a priceless tool for early heart disease prediction.

Advantages of Proposed System:

- Accurate Predictions
- User-friendly Interface
- Early Detection
- Continuous Improvement
- Data Privacy
- Enhanced public Health

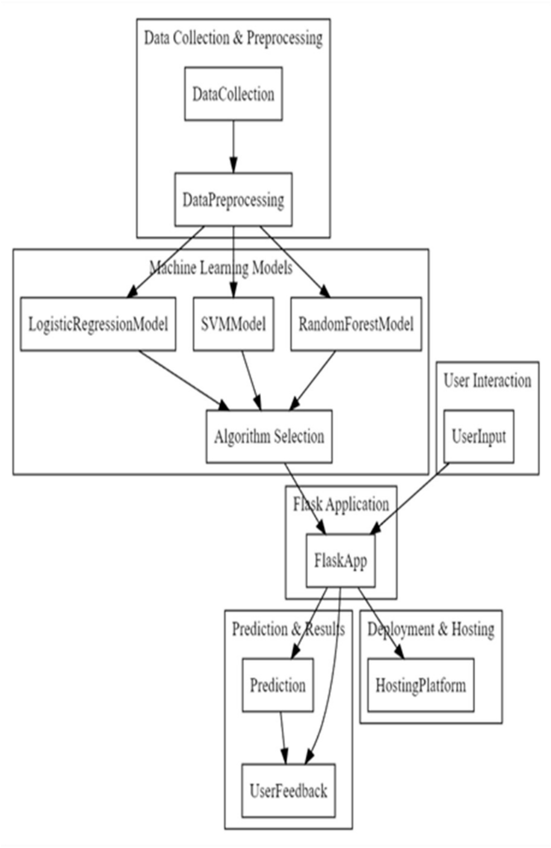

Fig:3 Proposed System Design

IV. MODEL EVALUATION

Model assessment is a critical step that involves assessing how well predictive analytics techniques perform on the testing sample. During the assessment process, a number of metrics are used to assess how effectively the models anticipate the existence or lack of cardiac illness.

1. Measures Utilized:

Various measures are used to thoroughly evaluate the models using the three ML techniques:

Accuracy (Correct Classification Rate): Determines the proportion of successfully predicted instances to all occurrences in order to assess the overall accuracy of predictions.

Precision (Positive Predictive Value-PPV): Measures how many real positive predictions there are among those that are projected to be positive. This gives information about how well the model avoids producing false positive results.

Recall (Sensitivity): Measures the sensitivity of the model to identify positive cases by split the overall number of real positive occurrences from the true positive forecasts.

***The F-Measure*:** The F-Measure estimates the harmonious average of both recall and accuracy in order to offer an honest evaluation of the effectiveness of a model in the classification of binary data.

AUC-ROC: The area underneath the destinator operational feature curve, or auc-roc, quantifies the relationship among the rate of real positives and the similarity of false positive (type I error). This is particularly crucial for imbalanced datasets.

In **Table. II** We have shown the sophisticated predictive analytics techniques' successful classification for identifying cardiac ailments. the Logistic Regression has the greatest categorization accuracy of 90% when choosing features approaches are used to foresee cardiovascular illness. Additionally, we discovered that sensitivity and reliability measurements show excellent performance. Below are the formulae for the measured factors:

$$\boldsymbol{CCR} = \frac{\mathrm{CP} + \mathrm{CN}}{CP + I + II + CN}$$

$$\boldsymbol{Positive\ Predictive\ Value} = \frac{\mathrm{CP}}{CP + I}$$

$$\boldsymbol{Sensitivity} = \frac{\mathrm{CP}}{CP + II}$$

$$\boldsymbol{F - Measure} = \frac{2 * (Sensitivity * PPV)}{Sensitivity + PPV}$$

CP = Correct Positive, ***CN*** = Correct Negative, ***I*** = False Type I Error, ***II***= False Type II Error.

A. Comparison of Result

The comparison includes a look at the log-linear classifier , maximum-margin classifier, and ensemble learning approach model. The system that consistently demonstrates the best quality of correct classification rate, positive predictive value, recall, F-Measure, as well as AUC-ROC scores emerges out in terms of cardiovascular illness diagnosis.

B. Algorithm Selection

- To get a thorough grasp of the learned systems' prediction abilities, a variety of regulations such as correct classification rate, positive predictive value , Sensitivity, and F-Measure, were carefully assessed.
- An algorithm selection module is essential for assessing and selecting the best predictive analytics techniques over accurate risk assessment of cardiovascular disease.
- This module guarantees data-driven choice making by calculating performance metrics and comparing log-linear classifier , maximum-margin classifier, and ensemble learning approach.

- The selection process considers • Correct classification rate, positive predictive value, sensitivity, and F-measure are taken into account throughout the selection process, enabling an educated but flexible response to modifications in the information set or computational performances.
- Finally, The Algorithm Selection involved a comprehensive comparison of model accuracies, leading to the identification of Logistic Regression as the most accurate algorithm, boasting an impressive accuracy rate of **90%**.

Therefore, across all three metrics (correct classification rate, positive predictive value , Sensitivity, and F-Measure), Logistic Regression demonstrated superior performance compared to Random Forest and Support Vector Machine. It appears to be the most accurate and well-balanced model for the heart disease prediction task in our project.

Table II: Performance Measure of Classifiers

Parameter Metrics	**Random Forest**	**SVM**	**Logistic Regression**
Accuracy	88%	66%	90%
PPV	84%	48%	88%
Sensitivity	84%	55%	85%
F-Measure	84%	51%	86%

C. Time Complexity of Logistic Regression with Graph

According to the dataset used in above algo, we have developed a graph of time complexity of Logistic Regression. From this graph we can compare Logistic Regression with other two approaches. And in this manner we can state that the time needed for the Logistic Regression algo, approach is smaller than the other two procedures.

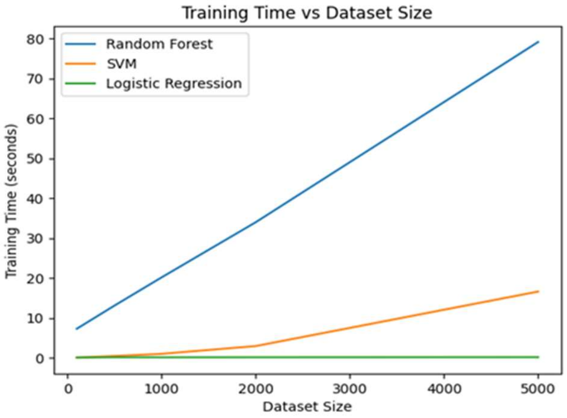

Fig:4 Time Complexity

At the end of our project using advanced predictive analytics techniques like log-linear classifier, maximum-margin classifier, and ensemble learning approach model, the heart disease prediction research project is able to provide predictions that are dependable and accurate. With a 90% accuracy rate, the selection process finds that the most efficient approach is logistic regression. The research, which is incorporated into a Flask-based web application, shows how to forecast cardiac illness early and effectively, so promoting proactive healthcare intervention.

V. FUTURE SCOPE

Future plans for our heart disease prediction research include adding more varied datasets, using cutting-edge machine learning methods, and improving the system's interpretability. More research into possible partnerships with medical facilities and the use of genetic factors in personalized predictions could both make the system work better and help cardiovascular health studies move forward.

VI. CONCLUSION

The overall objective of our work on heart disease prediction marks a noteworthy progression in healthcare technology by using machine learning to facilitate early detection and intervention. A thorough approach to predictive modelling is ensured by the combination log-linear classifier , maximum-margin classifier, and ensemble learning techniques. This one is the best method is log-linear classifier, which has an astounding 90% accuracy rate. Proactive health management is encouraged by the user-friendly interface of the Flask-based web application, which offers real-time forecasts. The project's flexibility, capacity to adjust to a variety of datasets, and dedication to data protection are key factors in its success. The potential future direction of the system is investigating genetic elements for individualized forecasts, working with healthcare institutions to extend data sources, and continuously improving using advanced machine learning methods. In addition to advancing the area of cardiovascular health research, this all-encompassing strategy opens the door for creative preventive healthcare solutions that eventually enhance public health outcomes. The project is a tribute to the capability of technology to bring about change in healthcare procedures and emphasizes the need of early identification in lessening the impact of cardiovascular illness both individuals or society.

REFERENCES

[1] S. Mondal, R. Maity, Y. Omo, S. Ghosh and A. Nag, "An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual-Stage Stacked Machine Learning Approaches," in IEEE Access, vol. 12, pp. 7255-7270, 2024, doi: 10.1109/ACCESS.2024.3350996.

[2] B. Ramesh and K. Lakshmanna, "A Novel Early Detection and Prevention of Coronary Heart Disease Framework Using Hybrid Deep Learning Model and Neural Fuzzy Inference

System," in IEEE Access, vol. 12, pp. 26683-26695, 2024, doi: 10.1109/ACCESS.2024.3366537.

[3] Z. Sun, W. Dong, J. Shi and Z. Huang, "Interpretable Disease Progression Prediction Based on Reinforcement Reasoning Over a Knowledge Graph," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 54, no. 3, pp. 1948-1959, March 2024, doi: 10.1109/TSMC.2023.3331847.

[4] W. Zhang, W. Cheng, K. Fujiwara, R. Evans and C. Zhu, "Predictive Modeling for Hospital Readmissions for Patients with Heart Disease: An updated review from 2012-2023," in IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2023.3349353.

[5] M. Bakro et al., "Building a Cloud-IDS by Hybrid Bio-Inspired Feature Selection Algorithms Along With Random Forest Model," in IEEE Access, vol. 12, pp. 8846-8874, 2024, doi: 10.1109/ACCESS.2024.3353055.

[6] A. Delilbasic, B. Le Saux, M. Riedel, K. Michielsen and G. Cavallaro, "A Single-Step Multiclass SVM Based on Quantum Annealing for Remote Sensing Data Classification," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 1434-1445, 2024, doi: 10.1109/JSTARS.2023.3336926.

[7] C. Xing, M. Wang, Y. Xu and Z. Wang, "SOML: Structure-Wise Ordinal Measure Learning for Hyperspectral Image Classification," in IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1-12, 2024, Art no. 5006912, doi: 10.1109/TIM.2023.3342836.

[8] M. Cheraghy, M. Soltanpour, H. B. Abdalla and A. H. Oveis, "SVM-based Factor Graph Design for Max-SR Problem of SCMA Networks," in IEEE Communications Letters, doi: 10.1109/LCOMM.2024.3366426.

[9] Y. Qiu, R. Li and X. Zhang, "Simultaneous SVM Parameters and Feature Selection Optimization Based on Improved Slime Mould Algorithm," in IEEE Access, vol. 12, pp. 18215-18236, 2024, doi: 10.1109/ACCESS.2024.3351943.

[10] H. B. Patel and N. J. Patil, "Enhanced CNN for Fruit Disease Detection and Grading Classification Using SSDAE-SVM for Postharvest Fruits," in IEEE Sensors Journal, vol. 24, no. 5, pp. 6719-6732, 1 March1, 2024, doi: 10.1109/JSEN.2023.3342833.

[11] A. Muhammed, P. Saji, M. J, V. V and S. K. P. R, "An Efficient Method to Localize and Quantify Axial Displacement in Transformer Winding Using Support Vector Machines," in IEEE Transactions on Industry Applications, vol. 60, no. 1, pp. 1827-1836, Jan.-Feb. 2024, doi: 10.1109/TIA.2023.3325210.

[12] Z. Liang and S. Ding, "Fuzzy Twin Support Vector Machines With Distribution Inputs," in IEEE Transactions on Fuzzy Systems, vol. 32, no. 1, pp. 240-254, Jan. 2024, doi: 10.1109/TFUZZ.2023.3296503.

[13] Y. Salini, S. N. Mohanty, J. V. N. Ramesh, M. Yang and M. M. V. Chalapathi, "Cardiotocography Data Analysis for Fetal Health Classification Using Machine Learning

Models," in IEEE Access, vol. 12, pp. 26005-26022, 2024, doi: 10.1109/ACCESS.2024.3364755.

[14] X. Lu, H. U. Sami and B. Güler, "SCALR: Communication-Efficient Secure Multi-Party Logistic Regression," in IEEE Transactions on Communications, vol. 72, no. 1, pp. 162-178, Jan. 2024, doi: 10.1109/TCOMM.2023.3308954.

[15] X. Lu, H. U. Sami and B. Güler, "SCALR: Communication-Efficient Secure Multi-Party Logistic Regression," in IEEE Transactions on Communications, vol. 72, no. 1, pp. 162-178, Jan. 2024, doi: 10.1109/TCOMM.2023.3308954.

[16] Gypsy Nandi, "Probabilistic Reasoning," in Principles of Soft Computing Using Python Programming: Learn How to Deploy Soft Computing Models in Real World Applications , IEEE, 2024, pp.159-196, doi: 10.1002/9781394173167.ch5.

[17] A. T. Olanipekun and D. Mashao, "HardnessTesterV: A Web Machine Learning application for Vickers Hardness Prediction of a Metallic Alloy Using Flask API," 2023 International Conference on Electrical, Communication and Computer Engineering (ICECCE), Dubai, United Arab Emirates, 2023, pp. 1-4, doi: 10.1109/ICECCE61019.2023.10442446.

[18] R. R, H. F. I, A. M, A. M. J and D. S, "Web Application Security Testing Framework using Flask," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 1646-1652, doi: 10.1109/ICAAIC56838.2023.10140422.

[19] C. I. Gunardi, A. Rashida, Hendrawan, E. Mulyana and W. Hermawan, "Web-Based Gender Classification ML Application Development for e-KYC," 2023 17th International Conference on Telecommunication Systems, Services, and Applications (TSSA), Lombok, Indonesia, 2023, pp. 1-5, doi: 10.1109/TSSA59948.2023.10366938.

[20] Shalabh Aggarwal, Flask Framework Cookbook: Enhance your Flask skills with advanced techniques and build dynamic, responsive web applications , Packt Publishing, 2023.