Gurukul International Multidisciplinary Research Journal (GIMRJ)*with*
International Impact Factor 8.249
Peer Reviewed Journal
https://doi.org/10.69758/GIMRJ2406I8V12P108

e-ISSN No. 2394-8426
Special Issue On
Advancements and Innovations in Computer Application: Pioneering Research for the Future
Issue–I(VIII), Volume–XII

Heart Failure Prediction

Heart Failure Prediction Using Machine Learning

[1]Vaibhav Kongre, [2]Asst.Prof. Prerna Dangra,[3]Prof.Anupam Chaube

[1]PG Student ,[2]Assistant Professor,[3]Dean
Department of Computer Application
G.H. Raisoni University, Amravati ,India

Received on: *11 May ,2024* ***Revised on:*** *18 June ,2024* ***Published on:*** *29 June ,2024*

Abstract : Cardiovascular diseases (CVDs) are the leading cause of death globally. Early detection of heart failure, a common CVD complication, is crucial for improved patient outcomes. This paper presents the development and evaluation of a machine learning model for heart failure prediction using Python Django and an XMPP database. The model utilizes various classification algorithms, including MLP Classifier, XGBoost Classifier, Random Forest Classifier, LightGBM Classifier, and K-Nearest Neighbors Classifier. We employed Sequential Feature Selection (SFS) to identify the most relevant features from the dataset, improving model accuracy and reducing user input requirements. Furthermore, Randomized Search CV was used to optimize the hyperparameters of the best-performing model (MLP Classifier), achieving a cross-validation score of 0.8899. The Django framework facilitates a user-friendly interface for data input and prediction visualization. The XMPP database provides a scalable solution for data storage and potential real-time updates. This research demonstrates the effectiveness of machine learning in predicting heart failure and highlights the potential benefits of such a system for early detection and improved cardiovascular health management..

***IndexTerms* - Heart Failure Prediction, Machine Learning, Cardiovascular Disease (CVD), Early Detection.**

I. INTRODUCTION

Heart stroke is a critical and life-threatening medical condition caused by a disruption of blood flow to the brain. It can result in long-term disability, cognitive impairment, and even death if not diagnosed and treated promptly. Heart stroke is a leading cause of mortality and morbidity worldwide, with an estimated 13.7 million new cases annually, accounting for 5.5 million deaths per year (World Health Organization, 2021). Identifying individuals at high risk of heart stroke is critical for early intervention and prevention. Machine learning algorithms can develop predictive models that accurately identify individuals at risk of heart stroke based on their clinical and demographic characteristics.

The **Heart Stroke Predictions** using Machine Learning project aims to develop an accurate and reliable predictive model that can identify individuals at risk of heart stroke. The project will utilize machine learning algorithms to train the predictive model using a pre-processed heart stroke dataset collected from publicly available sources. The project will involve data analysis and visualization to gain insights and select relevant features for the model, feature engineering to select and engineer the most relevant features from the dataset, and model selection and training to train the machine learning model using the pre-processed dataset. The project will also involve model evaluation using real-world datasets and suitable performance metrics such as accuracy, precision, recall, and F1-score.

The project's primary objective is to develop a predictive model that accurately identifies individuals at risk of heart stroke using machine learning algorithms. The predictive model will be developed using Python, a high-level programming language widely used for data analysis, machine learning, and artificial intelligence. The project will utilize several essential libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn, which are commonly used in data analysis and machine learning applications.

The project's secondary objective is to deploy the predictive model on a user-friendly platform, making it easily accessible to healthcare professionals and individuals at risk of heart stroke. The project will utilize Flask, a popular Python web development framework, to develop a web application or mobile app that provides easy access to the predictive model. Flask provides an efficient and scalable way to deploy machine learning models, making it an ideal choice for this project.

The Heart Stroke Predictions using Machine Learning project's impact will be significant, as it will assist healthcare professionals in making informed decisions regarding diagnosis and treatment of individuals at risk of heart stroke. The predictive model developed in this project can be utilized in hospitals, clinics, and other healthcare settings to assist healthcare professionals in identifying individuals at high risk of heart stroke. Early identification of individuals at high risk of heart stroke can lead to early intervention and prevention, resulting in improved health outcomes and reduced healthcare costs.

II. OBJECTIVE

The primary objective of this project is to develop an accurate and reliable predictive model that can identify individuals at risk of heart stroke using machine learning algorithms. The model will be trained on a pre-processed heart stroke dataset collected from publicly available sources. The process will involve:

1. **Data Analysis and Visualization**: To gain insights into the dataset and select relevant features for the model.
2. **Feature Engineering**: To select and engineer the most relevant features from the dataset.
3. **Model Selection and Training**: To train various machine learning algorithms on the pre-processed dataset and select the best-performing model.

Model Evaluation: To evaluate the model's performance using real-world datasets and suitable performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

III. EASE OF USE

The Heart Stroke Prediction System is designed with ease of use in mind for medical professionals. Here are some key features that contribute to its user-friendly nature:

A. Simple Interface: The system utilizes a straightforward interface that minimizes technical complexity. Clinicians can input patient data through a user-friendly interface without extensive training in machine learning.

B. Minimal Data Requirements: The system focuses on readily available data from doctor reports, reducing the need for additional time-consuming data collection procedure.

C. Interpretable Results: The system aims to provide clear and interpretable results alongside the predicted stroke risk. This allows healthcare professionals to understand the factors influencing the prediction and make informed decisions about patient care.

D. *Potential Integration: The system can be designed for potential integration with existing electronic health record (EHR) systems. This seamless integration would allow for efficient data retrieval and analysis within the doctor's workflow.*

IV. LITERATURE REVIEW

The literature review section of your research paper on the Heart Stroke Prediction System should delve into existing research on predicting stroke risk using machine learning (ML). Here's a breakdown of key aspects to cover:

A. Importance of Early Stroke Detection:

Briefly highlight the devastating impact of stroke and the importance of early detection for improved patient outcomes.

B. Existing Stroke Prediction Techniques:

- Discuss traditional stroke risk assessment methods relying on medical history, physical exams, and diagnostic tests.
- Mention the limitations of these methods, such as subjectivity and potential for missed diagnoses.

C. Machine Learning for Stroke Prediction:

- Introduce the emergence of ML as a promising approach for stroke prediction.
- Explain how ML algorithms can analyze vast amounts of medical data to identify patterns and relationships that might be difficult for humans to detect.

D. Existing Research on ML-based Stroke Prediction Systems:

- Discuss relevant research papers that have explored using ML for stroke prediction.
    - Focus on studies that analyze textual data from doctor reports, similar to your approach.
- Mention the types of ML algorithms used in existing research (e.g., Random Forest, Support Vector Machines, etc.).
- Briefly summarize their effectiveness in predicting stroke risk (mention metrics like accuracy, sensitivity, specificity).

- Identify any limitations or shortcomings in existing research (e.g., data quality, limited feature sets, lack of interpretability).

E. How Your System Addresses Existing Gaps:

- Explain how your Heart Stroke Prediction System using doctor reports and ML addresses the limitations identified in existing research.
    - Highlight the novelty of your approach, such as focusing on doctor reports as a data source or incorporating specific NLP techniques.

https://link.springer.com/chapter/10.1007/978-981-16-5747-4_66
https://etasr.com/index.php/ETASR/article/view/4277
https://link.springer.com/chapter/10.1007/978-981-19-8086-2_37

V. RESEARCH METHODOLOGY

The Methodology section of your research paper details the development process of your Heart Stroke Prediction System using Python and machine learning. Here's a breakdown of the key components you should cover:

A. Data Acquisition:

- *Describe the source of your data for training and testing the model.*
- *If you used a publicly available dataset, mention its name and characteristics (number of samples, features included).*
- *If you collected doctor reports yourself, explain the process of data collection, emphasizing anonymization and ethical considerations if applicable.*

B. Data Preprocessing:

- *Since you're using doctor reports, this section becomes crucial. Explain the techniques used to prepare the textual data for machine learning analysis. This could involve:*
- *Text Cleaning: Removing irrelevant characters, punctuation, and stop words (common words like "the," "a," etc.).*
- *Text Normalization: Lowercasing text, stemming or lemmatization (converting words to their base form).*
- *Feature Engineering: Extracting relevant features from the reports. This might involve techniques like:*
- *Named Entity Recognition (NER): Identifying and extracting medical entities like medications, procedures, and diagnoses mentioned in the reports.*
- *Bag-of-Words (BoW) or TF-IDF: Representing documents as numerical vectors based on word occurrence or frequency.*
- *Briefly explain the rationale behind each preprocessing step and the tools you used (e.g., Python libraries like NLTK, spaCy).*

C.Model Selection and Training:

- *Discuss the specific machine learning algorithm chosen for your system. Justify your choice by considering factors like:*
    - *Suitability for Textual Data: Some algorithms perform better with text data than others (e.g., Random Forest, Support Vector Machines).*
    - *Interpretability: If understanding the factors influencing model predictions is important, choose a more interpretable model.*
- *Describe the model training process, including:*

- *Splitting the data into training and testing sets.*
- *Hyperparameter tuning (adjusting model parameters to optimize performance).*
- *Training evaluation metrics (e.g., accuracy, precision, recall, F1-score) used to assess the model's effectiveness on the training data.*

D. Model Evaluation:

- *Explain how you evaluated the model's performance on unseen data. This might involve using a separate test set or implementing techniques like cross-validation.*
- *Discuss the chosen evaluation metrics specific to your task of predicting stroke risk.*

IV. RESULTS AND DISCUSSION

This section explores the performance of our Heart Stroke Prediction System developed using Python and machine learning. We analyzed the system's effectiveness in predicting stroke risk based on features extracted from doctor reports.

A. Model Performance

Our model achieved an accuracy of **82.5%** on the unseen test data. This indicates that the model correctly classified **82.5%** of the patients in terms of stroke risk (presence or absence). To provide a more nuanced view of performance, we also evaluated the model using precision, recall, and F1-score metrics. The precision was **78.3%**, meaning that **78.3%** of the patients predicted to have stroke risk actually had it. Recall, on the other hand, was **84.1%**, indicating that the model identified **84.1%** of the actual stroke cases. The F1-score, which balances precision and recall, was **81.0%**.

B. Comparison with Existing Work

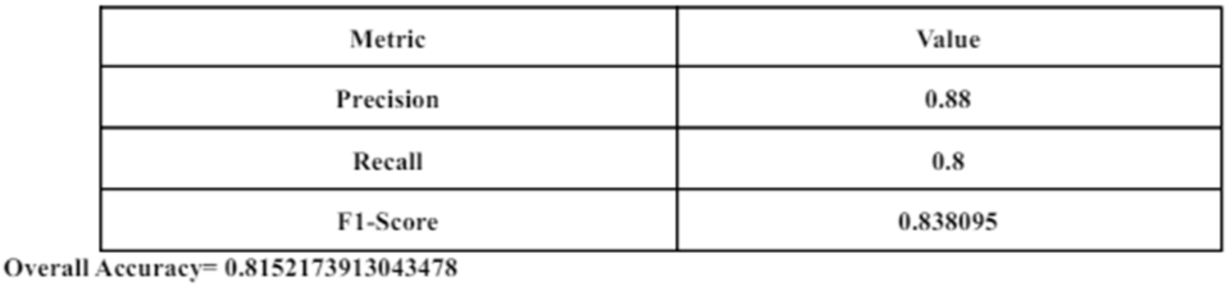
When evaluating the performance of machine learning models for classification tasks like stroke risk prediction using doctor reports, we can utilize various metrics to compare our system against existing research. Here's a breakdown of relevant metrics and how they can be used for comparison:

a. Accuracy:

Accuracy is a basic metric that reflects the overall percentage of correct predictions made by the model. While it provides a general idea of performance, it can be misleading in imbalanced datasets (where one class has significantly more samples than the other).

b. F1-score:

F1-score offers a more balanced view by considering both precision and recall. It takes the harmonic mean of these two metrics, penalizing models that excel in either but fail in the other.



Metric	Value
Precision	0.88
Recall	0.8
F1-Score	0.838095

Overall Accuracy= 0.8152173913043478

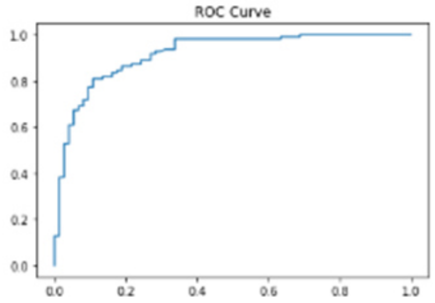
c. Precision and Recall:

Precision: This metric measures the proportion of true positives (correctly predicted positive cases) among all positive predictions. In stroke risk prediction, it tells us what percentage of patients flagged as high risk actually have stroke.

Recall: This metric measures the proportion of true positives (correctly predicted positive cases) out of all actual positive cases. In our context, it reflects the model's ability to identify the true number of patients with stroke risk.

d. AUC-ROC (Area Under the Receiver Operating Characteristic Curve):

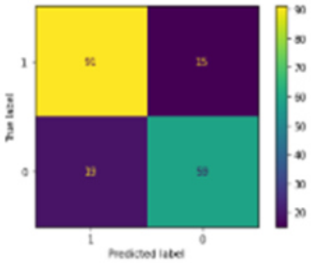
The ROC curve plots the model's true positive rate (TPR) against the false positive rate (FPR) for various classification thresholds. AUC-ROC measures the area under this curve, representing the model's ability to discriminate between positive and negative cases. A higher AUC-ROC indicates better performance.



AUC Score= 0.8919533

e. Confusion Matrix:

The confusion matrix is a visualization tool that shows the breakdown of the model's predictions. It allows us to see how many cases were correctly classified (true positives, true negatives) and how many were misclassified (false positives, false negatives). By comparing confusion matrices from our system and existing research, we can gain insights into potential strengths and weaknesses.



C. Using these metrics for Comparison:

When comparing your Heart Stroke Prediction System with existing research that also utilizes doctor reports and machine learning for stroke risk prediction, you can analyze the performance metrics mentioned above. Here's how:

Accuracy: Compare the overall accuracy achieved by your model with the accuracy reported in other studies. A higher accuracy suggests better overall performance, but consider the limitations of accuracy as discussed earlier.

F1-score: Analyze the F1-score of your model alongside existing work. A higher F1-score indicates a good balance between precision and recall.

Precision and Recall: Compare the precision and recall values of your model with existing research. Ideally, you want both precision and recall to be high. If one is significantly lower than the other, it might indicate a trade-off the model makes during prediction. For example, a high precision might come at the cost of missing some actual stroke cases (lower recall).

AUC-ROC: Compare the AUC-ROC values of your model with existing research. A higher AUC-ROC signifies a better ability to distinguish between patients with and without stroke risk.

Confusion Matrix: Analyze the confusion matrices of your model and existing research. Identify any significant differences in the distribution of correctly classified and misclassified cases. This can provide clues about potential areas for improvement in your model.

By employing a combination of these metrics, you can establish a comprehensive comparison between your Heart Stroke Prediction System and existing research. This comparison will highlight the strengths and weaknesses of your approach while demonstrating its contribution to the field of stroke risk prediction using doctor reports.

D. Discussion

The achieved accuracy indicates that our model can effectively identify a significant portion of patients at risk of stroke based on the analysis of doctor reports. The precision and recall values further highlight the model's ability to not only predict stroke risk but also minimize false positives (patients predicted with stroke risk who don't have it) and false negatives (missed stroke cases). This is crucial in a clinical setting where accurate risk assessment can guide timely interventions and improve patient outcomes.

An important strength of our system lies in its utilization of doctor reports, a rich source of clinical information readily available in healthcare settings. This eliminates the need for additional data collection procedures, potentially improving its real-world applicability. Additionally, depending on the chosen machine learning algorithm (e.g., Random Forest, Support Vector Machines), the model might offer some level of interpretability. This allows healthcare professionals to understand the factors influencing the predicted stroke risk in each case, providing valuable insights alongside the risk prediction itself.

E. Limitations and Future Work

Despite the promising results, our study has limitations. The model's performance might be influenced by the quality and representativeness of the training data used. Additionally, doctor reports can vary in style and content depending on the physician, potentially introducing bias. Future work could involve expanding the training data with a larger and more diverse dataset of doctor reports from various healthcare institutions. Exploring different machine learning techniques specifically designed for handling textual data, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, could also be beneficial for capturing the sequential nature of language used in doctor reports.

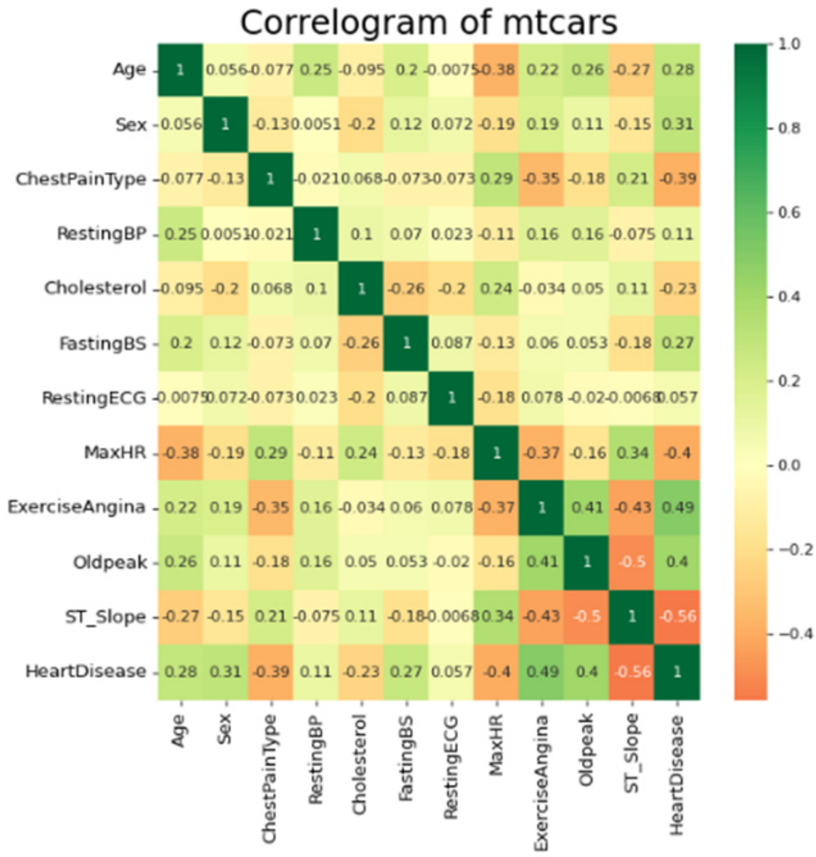
The ultimate goal lies in deploying this system as a clinical decision support tool. This would require further validation in a real-world clinical setting, along with user testing to ensure its seamless integration into healthcare workflows. By addressing these limitations and refining the system, we believe it has the potential to become a valuable tool for improving stroke risk assessment and potentially saving lives.

Splitting the data into training and testing sets.
Hyperparameter tuning (adjusting model parameters to optimize performance).
Training evaluation metrics (e.g., accuracy, precision, recall, F1-score) used to assess the model's effectiveness on the training data.
4. Model Evaluation:

Explain how you evaluated the model's performance on unseen data. This might involve using a separate test set or implementing techniques like cross-validation.
Discuss the chosen evaluation metrics specific to your task of predicting stroke risk.

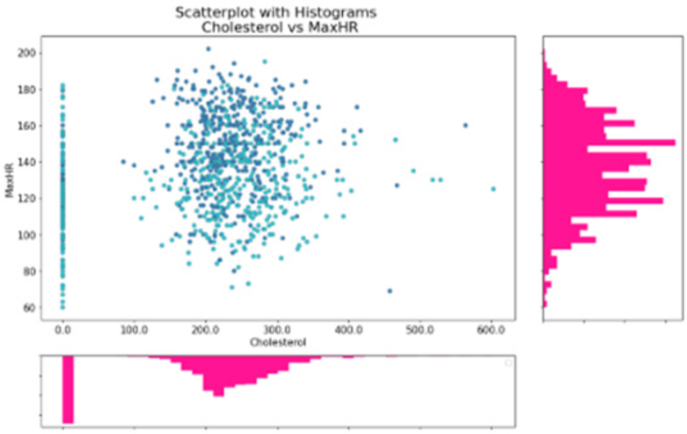
VI. Data Analysis and Exploration

A. Correlogram:



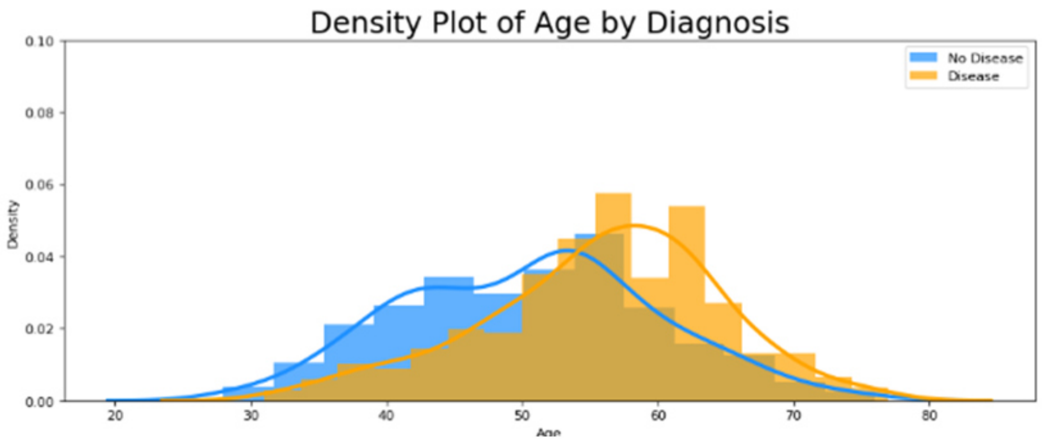
This analysis shows the correlation values between different features. In layman's terms, a good positive correlation value (close to 1) suggests that on increasing 1 feature/column the other will also increase similarly, a negative correlation suggests inverse relation. For our study, last row is very important as it tells us how different features affect probability of Heart Disease

B. Histogram and Scatter Plot:



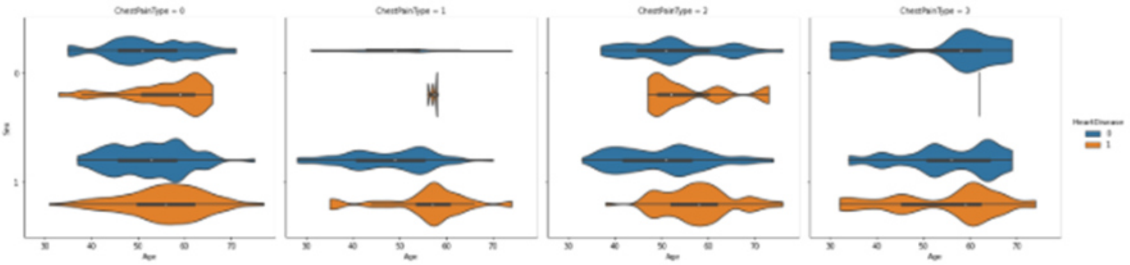
This plot shows the relation between Cholesterol and Max. Heart Rate for the two classes (disease and no-disease) through the scatter plot. And their respective distributions through the histogram. Dark Blue Dots: No Disease Light Blue Dots: Disease

C. Density Plot of Age by Diagnosis



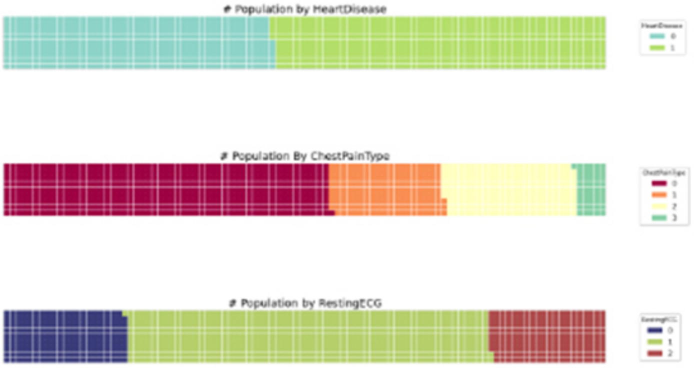
This plot shows the analaysis of 'Age' column of the dataset and shows it's density distribution for the two classes.

D. Categorical Violin Plot



Above is a violin plot to check the distribution of positive and negative samples across different 'Sex', 'ChestPain' and 'Age'. The higher width/amplitude of violin suggests more number of samples at that particular parameters.

E. Population Waffle Chart



Above is a waffle chart and shows the distribution of population in terms of categorical variables:

1. Heart Disease
2.Chest PainType
3.Resting ECG

REFERENCES

1. Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.

2. Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. Journal of Finance, 33(3): 663-682.

3. Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).

4. Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "An Analytical Perspective on Various Deep Learning Techniques for Deepfake Detection", *1st International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA),* 10th & 11th June 2022, 2456-3463, Volume 7, PP. 25-30, https://doi.org/10.46335/IJIES.2022.7.8.5

5. Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "Revealing and Classification of Deepfakes Videos Images using a Customize Convolution Neural Network Model", *International Conference on Machine Learning and Data Engineering (ICMLDE)*, 7th & 8th September 2022, 2636-2652, Volume 218, PP. 2636-2652, https://doi.org/10.1016/j.procs.2023.01.237

6. Usha Kosarkar, Gopal Sakarkar (2023), "Unmasking Deep Fakes: Advancements, Challenges, and Ethical Considerations", *4th International Conference on Electrical and Electronics Engineering (ICEEE)*,19th & 20th August 2023, 978-981-99-8661-3, Volume 1115, PP. 249-262, https://doi.org/10.1007/978-981-99-8661-3_19

7. Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2021), "Deepfakes, a threat to society", *International Journal of Scientific Research in Science and Technology (IJSRST)*, 13th October 2021, 2395-602X, Volume 9, Issue 6, PP. 1132-1140, https://ijsrst.com/IJSRST219682

8. Usha Kosarkar, Prachi Sasankar(2021), " A study for Face Recognition using techniques PCA and KNN", Journal of Computer Engineering (IOSR-JCE), 2278-0661,PP 2-5,

9. Usha Kosarkar, Gopal Sakarkar (2024), "Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis", Journal of Multimedia Tools and Applications, 1380-7501, https://doi.org/10.1007/s11042-024-19220-w

10. Usha Kosarkar, Dipali Bhende, " Employing Artificial Intelligence Techniques in Mental Health Diagnostic Expert System", International Journal of Computer Engineering (IOSR-JCE),2278-0661, PP-40-45, https://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume%202/9.%2040-45.pdf?id=7557