UCI Heart Disease

Ahmed Mohamed Ali[#1], Ali Sherif Badran[#2],

Omar Nabil[#3], Osama Mohamed Ali[#4]

Systems and Biomedical Engineering Department, Cairo University

[*] *AI in medical field*

***Abstract*— This paper presents a machine learning approach to predict heart disease using a comprehensive dataset that includes 14 attributes related to patient demographics, clinical symptoms, and diagnostic results. The dataset, sourced from the Cleveland heart disease database, encompasses diverse features such as age, sex, chest pain type, resting blood pressure, serum cholesterol, and more. The primary objective is to develop predictive models that can accurately determine the presence of heart disease based on these attributes. Various preprocessing steps were applied to handle missing values, scale numerical features, and encode categorical variables. We implemented several machine learning algorithms, including:LogisticRegression,DecisionTreeClassifier,RandomForest Classifier, VotingClassifier, and XGBClassifier. Additionally, ensemble learning using 50 random forest models and GridSearchCV were employed to optimize the hyperparameters of the random forest model. The best hyperparameters obtained from GridSearchCV were utilized in the VotingClassifier to enhance predictive performance. Our results indicate that the ensemble methods, particularly the optimized RandomForestClassifier and XGBClassifier, achieved the highest accuracy and robustness. This study underscores the potential of advanced machine learning techniques in enhancing early diagnosis and management of heart disease, providing a valuable tool for healthcare professionals.**

***Keywords*— Machine Learning, Heart Disease Prediction, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Voting Classifier, Ensemble Learning, GridSearchCV Hyperparameter Optimization, XGBoost**

I. Introduction

a. Problem Explanation

Heart disease is a leading cause of mortality worldwide, significantly impacting public health and healthcare systems. Early identification and accurate prediction are crucial for effective prevention and treatment, potentially saving lives and reducing healthcare costs. Despite advances in medical diagnostics, predicting heart disease remains challenging due to the complex interplay of various risk factors.

b. Motivation

The urgent need for reliable diagnostic tools drives this research. Machine learning (ML) offers promising techniques to analyze large datasets and uncover hidden patterns, potentially aiding early detection of heart disease. By leveraging ML, we aim to develop robust predictive models to improve patient outcomes and resource efficiency in healthcare.

c. Background

Heart disease includes conditions like coronary artery disease, arrhythmias, and heart failure. Traditional diagnosis involves patient history, physical exams, and tests like ECG and cholesterol levels. These methods, though effective, are time-consuming and require expert interpretation.

Advancements in ML have shown potential in automating and enhancing disease prediction accuracy. Studies indicate ML models can process complex datasets and provide accurate predictions to support early diagnosis and treatment planning.

d. Input & Output

Input: The project uses a multivariate dataset from the Cleveland heart disease database with 14 attributes: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, ECG results, maximum heart rate, exercise-induced angina, ST depression by exercise, slope of peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia.

Output: A predictive model that determines the presence or absence of heart disease based on these attributes, providing binary predictions to assist healthcare professionals in making informed diagnostic decisions.

II. Related work:

The detection and diagnosis of heart disease using machine learning have been the focus of extensive research over the past few years. Researchers have explored a variety of algorithms and data sets to enhance the accuracy and efficiency of predictive models in this critical field. This section reviews significant contributions from recent studies, highlighting the methodologies, data sets, and comparative performance of different machine learning approaches in heart disease detection.

The first study titled :" Comprehensive evaluation and performance analysis of machine learning in heart disease prediction " [1]

Dataset

The study utilizes the Cleveland heart disease dataset, which is obtained from the University of California, Irvine (UCI) online machine learning and data mining repository. The dataset contains 303 instances, but 6 of these have missing class values, leaving 297 instances for analysis. It includes 76 attributes per subject, but only 13 are used for the detection of heart disease. These attributes are a mix of categorical and numerical features, such as age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, etc. .

Algorithms Used

The primary algorithm used in the study is the XGBoost (Extreme Gradient Boosting) classifier. The model's hyperparameters were optimized using Bayesian optimization. The study also compares

the performance of the XGBoost model with two other machine learning models:

- Random Forest (RF)
- Extra Tree (ET)

Accuracy Comparison Between Algorithms

- Proposed XGBoost Model: 91.80%
- Random Forest (RF) Model: 88.52%
- Extra Tree (ET) Model: 88.52%

Detailed Accuracy Comparison

The accuracy of the proposed XGBoost model is higher than both the Random Forest and Extra Tree models. The XGBoost model achieved an accuracy of 91.80%, while both the Random Forest and Extra Tree models achieved an accuracy of 88.52%. The XGBoost model, optimized with Bayesian optimization and using One-Hot encoding for categorical features, shows superior performance in predicting heart disease compared to the Random Forest and Extra Tree models.
This shows the strength of the XGBoost algorithm and how it can be a strong candidate to solve this issue.

The second paper titled: "Heart Disease Diagnosis Using Machine Learning Algorithms" [2]

Data Set:

The study utilizes the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. This dataset consists of 303 instances with 76 attributes, of which only 14 are used in the analysis. The attributes include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, old peak (ST depression induced by exercise), slope of the peak exercise ST segment, number of major vessels (0-3) colored by fluoroscopy, and thalassemia.

Algorithms Used:

- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Naïve Bayes (NB)
- Decision Tree (DT)
- Random Forest (RF)
- Logistic Regression (LR)
- Artificial Neural Network (ANN).

Accuracy Comparison Between Algorithms:

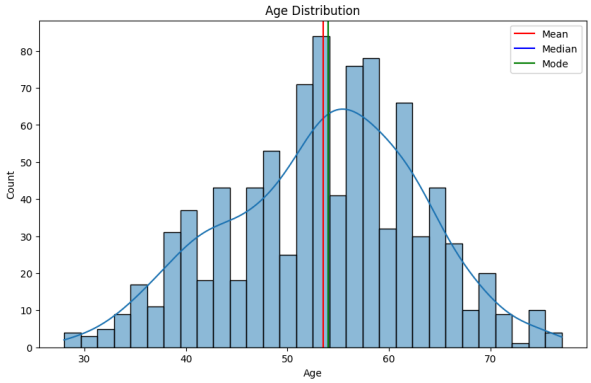
Accuracy Results:

- SVM: 84.1%
- KNN: 82.0%
- NB: 83.6%
- DT: 78.0%
- RF: 85.3%
- LR: 82.5%
- ANN: 85.0%

The Random Forest algorithm achieved the highest accuracy (85.3%), closely followed by the Artificial Neural Network (85.0%) and Support Vector Machine (84.1%). The Decision Tree had the lowest accuracy among the evaluated algorithms.

III. Dataset and Features

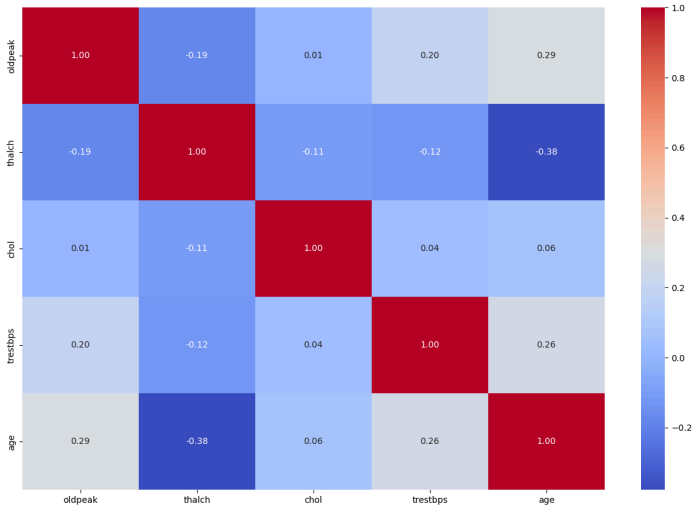
1. id (Unique id for each patient): This feature represents a unique identifier for each patient in the dataset.
2. age (Age of the patient in years): The age of the patient in years, ranging from a minimum of 28 to a maximum of 77. The mean age is approximately 53.5 years, with a median of 54 years.



3. origin (Place of study): Indicates the origin or location of the study, with data collected from multiple locations including Cleveland, Hungary, VA Long Beach, and Switzerland.
4. sex (Male/Female): Represents the gender of the patient, with male subjects comprising approximately 78.91% of the dataset, while female subjects make up about 21.09%.
5. cp (Chest Pain Type): Describes the type of chest pain experienced by patients, categorized into four types: typical angina, atypical angina, non-anginal pain, and asymptomatic.
6. trestbps (Resting Blood Pressure): Represents the resting blood pressure of patients measured in mm Hg on admission to the hospital. The average resting blood pressure is approximately 132 mm Hg.
7. chol (Serum Cholesterol): Indicates the serum cholesterol levels of patients measured in mg/dl, with an average serum cholesterol level of approximately 199 mg/dl.
8. fbs (Fasting Blood Sugar): Represents whether the fasting blood sugar level of patients is greater than 120 mg/dl (1 = true, 0 = false).
9. restecg (Resting Electrocardiographic Results): Describes the results of resting electrocardiography, categorized into three values: normal, ST-T wave abnormality, and left ventricular hypertrophy.
10. thalach (Maximum Heart Rate Achieved): Indicates the maximum heart rate achieved during exercise testing.
11. exang (Exercise-induced Angina): Represents whether patients experience exercise-induced angina (1 = true, 0 = false).
12. oldpeak (ST Depression Induced by Exercise Relative to Rest): Quantifies the magnitude of ST depression induced by exercise relative to rest.
13. slope (Slope of the Peak Exercise ST Segment): Describes the slope of the peak exercise ST segment, categorized into three values: upsloping, flat, and downsloping.
14. ca (Number of Major Vessels Colored by Fluoroscopy): Represents the number of major vessels colored by fluoroscopy, ranging from 0 to 3.

15. thal (Thalassemia): Describes the thalassemia condition of patients, categorized into three values: normal, fixed defect, and reversible defect.
16. num (The Predicted Attribute): Represents the predicted attribute indicating the presence or absence of heart disease, with values ranging from 0 to 4.

The correlation matrix showed no significant correlations between the features, indicating that they operate independently of each other.



1. Outliers

   The row with a resting blood pressure (trestbps) value of 0 was removed from the dataset as it was deemed to be an error in the data.

2. Missing Values

   Missing values were present in several columns, with some exhibiting higher ratios of missing values.

   Advanced imputation techniques were applied to address missing values, including the use of iterative imputation with Random Forest Classifier and Random Forest Regressor for categorical and numerical columns, respectively.

   Iterative imputation was performed iteratively on each column with missing values, utilizing Random Forest models to predict and impute missing values.

   Missing values were imputed using a combination of label encoding for categorical variables and iterative imputation with Random Forest models for both categorical and numerical variables.

3. Consistency in Attribute Names

   Attribute names were standardized for consistency, with spaces removed and boolean values encoded.

4. Encoding Categorical Variables

   Categorical variables were encoded to numeric format, including encoding of boolean values for fasting blood sugar (fbs) and exercise-induced angina (exang).

5. Training, validation and testing

   We conducted both cross-validation and traditional train-test split to evaluate the performance of our models. For the train-test split, we allocated 70% of the dataset for training and 30% for testing. This partitioning ensured that our models were trained on a sufficiently large portion of the data while retaining an independent subset for evaluation. Additionally, we employed k-fold cross-validation with (k = 5) folds to further validate the robustness of our models. This technique involved dividing the data into (k) subsets, training the model on (k-1) subsets, and evaluating its performance on the remaining subset iteratively. By averaging the results over these iterations, we obtained a more reliable estimate of the model's performance and mitigated the risk of overfitting. Combining these methods allowed us to assess the generalization capabilities of our models effectively.

6. Final Dataset

   A new dataset was generated with a subset of necessary columns for the analysis, including age, sex, chest pain type, country of origin, various physiological attributes, and the target variable indicating the presence or absence of heart disease.

sample of dataset

1. Age: 63, Sex: Male, Dataset: Cleveland, CP: Typical Angina, Trestbps: 145, Chol: 233, FBS: True, Restecg: LV Hypertrophy, Thalch: 150.
2. Age: 67, Sex: Male, Dataset: Cleveland, CP: Asymptomatic, Trestbps: 160, Chol: 286, FBS: False, Restecg: LV Hypertrophy, Thalch: 108.
3. Age: 67, Sex: Male, Dataset: Cleveland, CP: Asymptomatic, Trestbps: 120, Chol: 229, FBS: False, Restecg: LV Hypertrophy, Thalch: 129.
4. Age: 37, Sex: Male, Dataset: Cleveland, CP: Non-Anginal, Trestbps: 130, Chol: 250, FBS: False, Restecg: Normal, Thalch: 187.
5. Age: 41, Sex: Female, Dataset: Cleveland, CP: Atypical Angina, Trestbps: 130, Chol: 204, FBS: False, Restecg: LV Hypertrophy, Thalch: 172.
6. Age: 56, Sex: Male, Dataset: Cleveland, CP: Atypical Angina, Trestbps: 120, Chol: 236, FBS: False, Restecg: Normal, Thalch: 178.

IV. Methodology:

We have implemented and compared several machine learning models for heart disease classification, including Logistic Regression, Naive Bayes, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest, and XGBoost. Each algorithm's specific characteristics and functionalities are described below. We used cross-validation to optimize the hyperparameters for all models except for Naive Bayes.

Logistic Regression

Logistic Regression is a linear model used for binary classification problems. It estimates the probability that a given

instance belongs to a particular class using the logistic function. The model predicts the probability p of the positive class (heart disease presence) as:

$$p = \sigma(\mathbf{w}^T\mathbf{x} + b) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

where σ is the sigmoid function, w is the vector of weights, x is the feature vector, and b is the bias term. The loss function used is the binary cross-entropy loss:

$$L(\mathbf{w}, b) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right]$$

where y_i is the true label and p_i is the predicted probability for the i-th instance.

Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' Theorem with the assumption of independence between features. We used the Gaussian Naive Bayes variant for this study, suitable for continuous data. The probability of a class C_K given a feature vector x is calculated as:

$$P(C_k|\mathbf{x}) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{P(\mathbf{x})}$$

Here,$P(x_i)|C_K$ follows a normal distribution:

$$P(x_i|C_k) = \frac{1}{\sqrt{2\pi\sigma_{C_k}^2}}\exp\left(-\frac{(x_i-\mu_{C_k})^2}{2\sigma_{C_k}^2}\right)$$

where $\boldsymbol{\mu}C_K$ and $\boldsymbol{\sigma}C_K$ are the mean and variance of the feature x_i in class C_K.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm used for classification. It classifies a data point based on how its neighbors are classified. Given a positive integer k and a test instance x, the algorithm performs the following steps:

1. Compute the distance (usually Euclidean) between x and all the training points.
2. Select the k training points that are closest to x.
3. Assign x to the class most common among the k-nearest neighbors.

The Euclidean distance between two points x_i and x_j is given by:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{m=1}^{M}(x_{i,m} - x_{j,m})^2}$$

where M is the number of features.

Support Vector Machine (SVM)

SVM is a supervised learning algorithm used for classification by finding the hyperplane that best separates the classes. For a feature vector x, the decision function is:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

The goal is to maximize the margin 2/||w|| between the hyperplane and the nearest data points (support vectors). This is formulated as a convex optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))$$

where C is a regularization parameter.

Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification. Each tree is built using a random subset of the training data and features. The algorithm reduces overfitting and improves generalization. The prediction for an instance x is:

$$\hat{y} = \text{mode}(\{h_t(\mathbf{x})\}_{t=1}^{T})$$

where h_t is the t-th decision tree and T is the total number of trees.

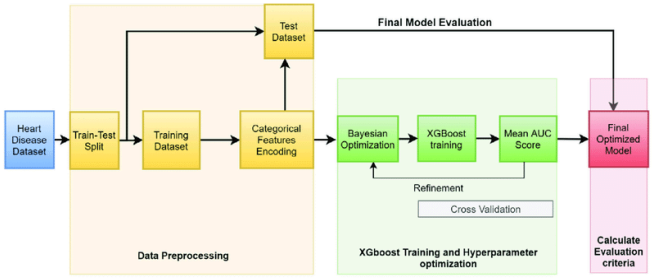
XGBoost

XGBoost is an optimized gradient boosting algorithm that builds an ensemble of weak learners, typically decision trees, in a sequential manner. It minimizes the following objective function:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K}\Omega(f_k)$$

where $\mathcal{L}$ is a differentiable convex loss function that measures the difference between the true label y_i and the prediction $\hat{y}_i$, and Ω is a regularization term that penalizes the complexity of the model to prevent overfitting.

The following is the block diagram followed for our dataset for typically using XGboost as we found it giving the best output metrics.



VI. Experiments/Results/Discussion:

Hyperparameter Selection

For all models except Naive Bayes, we used cross-validation to find the optimal hyperparameters. Cross-validation involves dividing the dataset into k subsets and training the model k times, each time using a different subset as the validation set and the remaining data for training. This method helps in assessing the model's performance and ensures that it generalizes well to unseen data. The optimal hyperparameters were chosen based on the average performance metric (e.g., accuracy, F1-score) across all folds. For each model, hyperparameters were tuned to optimize performance. The XGBoost model, for instance, had its parameters such as `learning_rate`, `n_estimators`, `max_depth`, `colsample_bytree`, `min_child_weight` and `subsample` optimized using cross-validation. Random Forest hyperparameters like `n_estimators`, `min_samples_leaf`, `min_samples_split` and max_depth were selected via cross-validation as well. Similarly, the SVM's regularization parameter `C`, `gamma` and `kernel` type were fine-tuned using a cross-validation. The KNN model's k value, `metric` and

`weights` were determined through cross-validation, while Naive Bayes did not require parameter tuning. As for the Logistic Regression model it had parameters like `C`, `solver` were optimized using cross-validation.

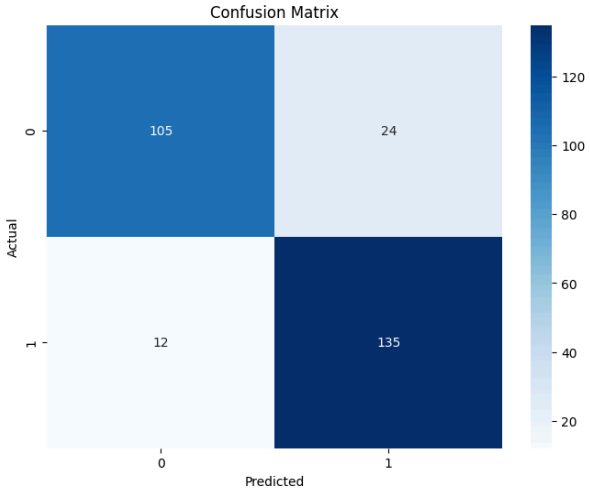
Primary Metrics

We evaluated our models using several key metrics:

- Accuracy: $Accuracy = (TP + TN)/(TP + FP + FN + TN)$
- Precision: $Precision = TP/(FP + TP)$
- Recall: $Recall = TP/(TP + FN)$
- F1-Score: $F1\ Score = 2\cdot(Precision\cdot Recall)/(Precision + Reca$
- Confusion Matrix: To illustrate the performance in terms of TP, FP, TN, and FN.

Classification Metrics

Below is the confusion matrix for the XGBoost model:



Quantitative and Qualitative Results

The table below summarizes the performance metrics of the evaluated models:

TABLE I

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	86.96%	0.87	0.86	0.86
Random Forest	83.70%	0.83	0.83	0.83
SVM	82.97%	0.82	0.82	0.82
KNN	84.42%	0.84	0.84	0.84
Naive Bayes	83.33%	0.83	0.83	0.83
Logistic Reg.	84.42%	0.84	0.84	0.84

VII. Conclusions:

We presented a comprehensive evaluation of various machine learning algorithms for predicting heart disease using the Cleveland heart disease dataset. The key algorithms assessed include XGBoost, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Logistic Regression.

Key Findings

XGBoost: Best performer with 86.96% accuracy, strong precision-recall balance, handles feature interactions well, and avoids overfitting through regularization.

Random Forest: 83.70% accuracy, high training accuracy but slightly overfits, leverages multiple decision trees for enhanced stability.

SVM: 82.97% accuracy, excels at finding optimal hyperplanes for classification.

KNN: 84.42% accuracy, effective in capturing local patterns, simple yet powerful.

Naive Bayes: 83.33% accuracy, efficient, balances bias and variance.

Logistic Regression: 84.42% accuracy, straightforward, provides good bias-variance trade-off.

Future Work

Given additional time, resources, and computational capabilities, several avenues could be explored to enhance the predictive models further:

1. Deep Learning: Investigate CNNs and RNNs for capturing complex patterns and temporal relationships.
2. Feature Engineering: Develop sophisticated techniques to capture non-linear relationships.
3. Model Interpretability: Use SHAP values to align predictions with medical expertise.
4. Class Imbalance Handling: Apply methods like SMOTE or cost-sensitive learning to improve performance on minority classes.
5. Additional Data Integration: Incorporate genetic, lifestyle, and longitudinal health data for a more comprehensive model.

References

[1] A. M. Al-Milli, "Backpropagation Neural Network for Prediction of Heart Disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 28, no.1,pp.40-44,Jan.2016.Available:https://doi.org/10.1016/j.jksuci.2014.02.002.

[2] S. K. Lakshmanaprabu, K. Shankar, A. L. Fine, P. Geman, and A. H. Eliason, "A Machine Learning-Based Framework for Heart Disease Detection," World Journal of Engineering and Technology, vol. 6, no. 4, pp. 444-456, Dec.2018.Available:https://www.scirp.org/journal/paperinformation.aspx?paperid=89127.

[3] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease Data," Hungarian Institute of Cardiology, Budapest; University Hospital, Zurich, Switzerland; University Hospital, Basel, Switzerland; V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. [Online].Available:https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/data. [Accessed: May 25, 2024].

[4] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.

[5] David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."

[6] Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11--61

TABLE II
TEAM CONTRIBUTIONS

Member name	**Section and Bench No.**	**Contributions**
Ali Sherif Badran	sec.1 B.N.40	Research on similar Papers EDA Modelling
Ahmed Mohamed Ali	sec.1 B.N.5	EDA Preprocessing Paper Formatting
Omar Nabil Mahmoud	sec.1 B.N.48	Research on similar Papers Preprocessing Modelling
Osama Mohamed Ali	sec.1 B.N.8	EDA Modelling Paper Formatting