HEART DISEASE IDENTIFICATION METHOD USING MACHINE-LEARNING CLASSIFICATION IN E-HEALTHCARE

Dr. Anita Mahajan[1], Ms.Krushna Jagtap[2*]

[1]*Assistant Professor, Department of Computer Engineering, Ajeenkya D.Y.Patil School Of Engineering, Lohegaon*

[2*]*Student Master Of Computer Engineering , Ajeenkya D.Y.Patil School Of Engineering ,Lohegaon*

Abstract: heart disease is a life-threatening disorder that needs prompt and correct diagnosis to enhance patient outcomes. Traditional diagnostic approaches often have limits in both accuracy and efficiency. As a result, numerous machine learning and deep learning algorithms have been used to automate the prediction of cardiac disease, delivering a greater degree of accuracy. This study uses a variety of algorithms, including Support Vector Machine (SVM), Naive Bayes, Decision Tree, Random Forest, and Artificial Neural Networks (ANN), to predict the presence or absence of heart disease based on patient characteristics such as age, gender, cholesterol levels, blood pressure, and other health indicators. Each method has distinct capabilities; for example, the Random Forest algorithm excels at managing complicated datasets by generating an ensemble of decision trees, while SVM is well-known for its capacity to handle high-dimensional data. Naive Bayes is simple and fast, making it ideal for real-time predictions. Meanwhile, Decision Trees are easily interpretable. Furthermore, ANNs are used to capture complicated, nonlinear connections in data, whereas Convolutional Neural Networks (CNNs) are investigated for deep learning-based feature extraction and classification. The findings show that integrating these models greatly improves the prediction accuracy and reliability of heart disease diagnosis.

Keywords: Heart disease, Diagnosis, Machine learning, Deep learning, Predictive modeling, Random Forest classifier, Feature extraction

1. Introduction

Heart disease encompasses a variety of disorders that impact the structure and function of the heart, such as arrhythmias, heart failure, and coronary artery disease. Effective cardiac disease prediction and timely diagnosis are essential for enhancing patient care, lowering medical expenses, and saving lives. While clinical signs and risk factor analysis are the foundation of traditional risk assessment approaches, machine learning techniques may improve prediction accuracy and help healthcare make well-informed choices.

This study compares and evaluates the effectiveness of several machine learning algorithms for the prediction of heart disease. Based on a set of clinical and demographic characteristics, the research seeks to determine the best model for correctly classifying individuals as having or not having heart disease. K-Nearest Neighbors, Random Forest, Neural Network, Decision Tree, and Logistic Regression are among the models that have been assessed.

Accuracy, precision, recall, and F1-score are among the relevant performance measures used in the training and assessment of the machine learning models. The study compares and contrasts the various machine learning models, looking at their prediction power, shortcomings, and strengths. The research also looks at how model performance is affected by data preprocessing methods such feature scaling and class balancing. The results reveal the most accurate model for predicting heart disease and provide insights into the efficacy of each algorithm. The results of this study may help academics and healthcare professionals choose the best machine learning strategy for predicting heart disease. Accurately identifying those who are at risk allows for the implementation of preventive interventions, which enhance patient outcomes and the distribution of resources within healthcare systems.

2. Problem Statement

The problem addressed in this research is the need for robust and accurate prediction models for heart disease using machine learning techniques. Traditional risk assessment methods have limitations, and there is a demand for improved early detection and risk stratification.

3. Objectives

3.1. To analyze and compare multiple machine learning and deep learning algorithms for the accurate prediction of heart disease based on key patient health indicators.

3.2. To develop a predictive model that can automatically classify individuals as having heart disease or not, using medical parameters such as age, gender, cholesterol level, blood pressure, and other relevant features.

3.3. To implement and evaluate different machine learning classifiers, including Support Vector Machine (SVM), Naive Bayes, Decision Tree, Random Forest, and Artificial Neural Networks (ANN), based on their prediction performance.

3.4. To explore the application of Convolutional Neural Networks (CNNs) for deep learning-based feature extraction and classification within structured medical data.

3.5. To enhance the accuracy and reliability of heart disease prediction by integrating the strengths of various algorithms through ensemble or hybrid modeling techniques.

3.6. To assess model performance using standard evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

3.7. To design a scalable and efficient e-healthcare decision support system that assists medical professionals in early detection and timely intervention for heart disease patients.

3.8. To ensure real-time or near-real-time prediction capabilities, especially using lightweight models like Naive Bayes for use in mobile or cloud-based e-health applications.

4. Risks and Benefits

4.1 Risks:

1. **Data Privacy Concerns:** Handling sensitive patient data poses a risk of breaches if not managed securely.
2. **Model Bias:** Imbalanced datasets may result in biased predictions, especially toward underrepresented groups.
3. **Overfitting:** Complex models like ANN or Random Forest can overfit if not properly tuned or validated.
4. **Misclassification:** Incorrect predictions could lead to false diagnoses, affecting patient treatment plans.
5. **System Reliability:** Dependence on algorithmic systems might reduce human oversight in critical cases.
6. **Hardware/Software Limitations:** Running deep learning models may require high-performance systems, which could be a challenge in low-resource environments.

4.2 Benefits:

1. **Improved Accuracy**: ML models significantly enhance diagnostic accuracy compared to traditional methods.
2. **Faster Diagnosis**: Automated systems provide rapid predictions, essential for timely medical intervention.
3. **Cost-Effective**: Reduces the need for expensive or invasive diagnostic procedures.
4. **Scalability**: Can be deployed across various healthcare settings, from hospitals to remote clinics.
5. **Real-Time Monitoring**: Enables integration with IoT devices for continuous health monitoring.
6. **Decision Support**: Assists medical professionals with evidence-based insights, supporting better treatment plans.
7. **Early Detection**: Helps identify potential heart issues before symptoms worsen, potentially saving lives.

5. Literature Survey

Dwarakanath B et.al (2022) In the context of e-healthcare, online illness detection services have become more popular and have increased in quality due to developments in data mining, wearable technology, and cloud computing. Through early illness detection, e-healthcare services contribute to a lower mortality rate. At the same time, heart disease (HD) is a fatal condition, and early HD diagnosis is essential for patient survival.

Santhosh V et.al (2022) Using information gathered from prior patients as well as information entered by the user at that specific moment, this project started with the early identification of all likely symptoms and indicators that might eventually lead to the discovery of cardiac illnesses. Health care data utilized for surveillance in the modern day is more than just a time-building sequence of daily counts. Rather, a multitude of suggested demographic and symptom data, both regional and temporal, are accessible at the moment of execution.

Mohan Raja. Pulicharla et.al (2023) The pursuit of artificial intelligence gave rise to the science of machine learning. In the early days of artificial intelligence as a subject of study, several scholars were fascinated by the concept of allowing computers to learn from data. In addition to what were then called "neural networks," which were essentially perceptrons and other models that were later shown to be reimaginings of the generalized linear models of statistics, they attempted to address the problem using a number of symbolic approaches. Additionally, probabilistic reasoning was used, especially for automated medical diagnosis.

Deepika Tenepalli et.al (2024) Since its use is necessary for day-to-day living, the Internet of Things (IoT) has been embraced in numerous applications in recent years. Additionally, the healthcare system is using this emerging technology to provide patients efficient emergency treatments. The number of illnesses and medical issues among individuals is rapidly increasing in the present situation. As a result, the healthcare industry began integrating IoT and assistive technologies to provide effective wireless healthcare services and to continuously monitor patients.

S Siamala Devi et.al (2021) One of the most unexpected illnesses is heart disease, which has affected many people worldwide. Accurate and timely evidence of coronary heart disease is essential to medical treatment, particularly in the field of cardiology. Based on AI processes, a useful and accurate paradigm for identifying cardiac disease is put forward. Medical data with 335 features and 26 features is used to test this method. The greatest notable expectation precision of 88.4% is achieved by MLR. This approach is also compared to data on coronary heart disease in Cleveland. Furthermore, in this case, MLR outperforms other techniques

Arokia Jesu Prabhu L et.al (2020) While other industries progress with the help of cognitive computing, the health care industry is still developing and providing more benefits to all customers. The aging population leads to poor decision-making, which has a negative influence on treatment quality and increases treatment costs, adding to the already complicated healthcare system. However, there are a number of obstacles that hinder progress in this area, including query inconsistencies, user domain information sets, and gaps between the knowledge base and user queries. From 1-D cardiovascular beatings to automated discovery utilizing multi-dimensional clinical data, the fast advancement of machine learning and artificial intelligence for medical applications has already been shown in recent years

6. Proposed System

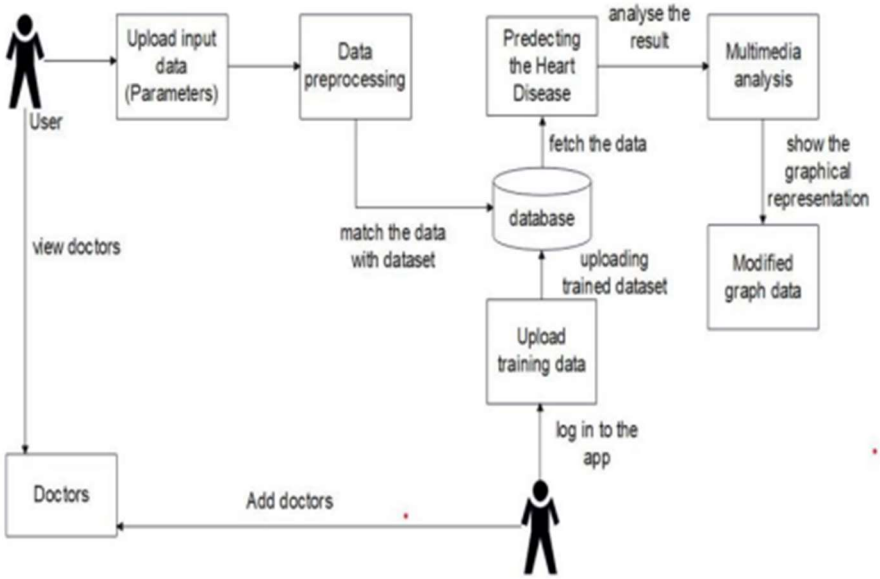


Fig:Sysytem Architecture

The primary objective of this research is to develop a robust machine learning-based model capable of predicting heart disease based on clinical and demographic data. The methodology is divided into several key stages: data collection, data preprocessing, model development, and model evaluation. Below, we provide a detailed breakdown of each phase, followed by a summary of the performance evaluation results for the various models tested.

1. Data Collection:

The dataset used for this research includes 8,763 patient records, each containing clinical data and lifestyle factors pertinent to heart disease. The dataset consists of 14 features, including both demographic variables (e.g., age, sex) and clinical measurements (e.g., cholesterol level, blood pressure, heart rate, stress level). Additionally, lifestyle-related factors such as smoking, alcohol consumption, family history, and diet are included. The target variable, *Heart Attack Risk*, is binary, indicating whether a patient is at high risk (1) or low risk (0) for a heart attack.

2. Data Preprocessing:

Before applying machine learning models, various preprocessing steps were performed to prepare the data:

- **Handling Missing Data**: Missing values were either imputed with appropriate methods or removed to ensure that the dataset was complete.
- **Encoding Categorical Data**: Categorical variables such as *Sex* and *Diet* were transformed into numerical values using one-hot encoding, which allows machine learning models to work with them efficiently.
- **Feature Scaling**: Numerical features, including *Cholesterol*, *Heart Rate*, and *Stress Level*, were standardized to ensure they all fall within a similar range, improving the performance of certain algorithms.
- **Data Splitting**: The dataset was divided into two subsets: 80% of the data was used for training the models, while the remaining 20% was reserved for testing and evaluating the model's performance.

3. Model Selection:

Several machine learning algorithms were considered for heart disease prediction, each offering different strengths and weaknesses. The models tested were:

- **Logistic Regression (LR)**
- **Naive Bayes (NB)**
- **Support Vector Machine (SVM)**
- **Decision Tree (DT)**
- **Random Forest (RF)**
- **Artificial Neural Networks (ANN)**

Each algorithm was trained on the training set, and its performance was evaluated using the test set to determine how effectively it predicts heart disease risk.

4. Model Evaluation

To assess the performance of each model, the following evaluation metrics were employed:

- **Accuracy**: The proportion of correct predictions made by the model relative to the total number of predictions.
- **Precision**: The ratio of correctly predicted positive cases to the total predicted positive cases, indicating how well the model identifies positive instances.
- **Recall**: The ratio of correctly predicted positive cases to the total actual positive cases, reflecting the model's ability to identify all relevant positive instances.
- **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two. This metric is particularly useful in dealing with imbalanced datasets, where one class is more prevalent than the other.

7. Result of Proposed System

Algorithm	Accuracy	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)
Logistic Regression	0.635	0.64	1	0.78
Naive Bayes	0.635	0.64	1	0.78
SVM	0.635	0.64	1	0.78
Decision Tree	0.537	0.64	0.63	0.63
Random Forest	0.623	0.64	0.95	0.76

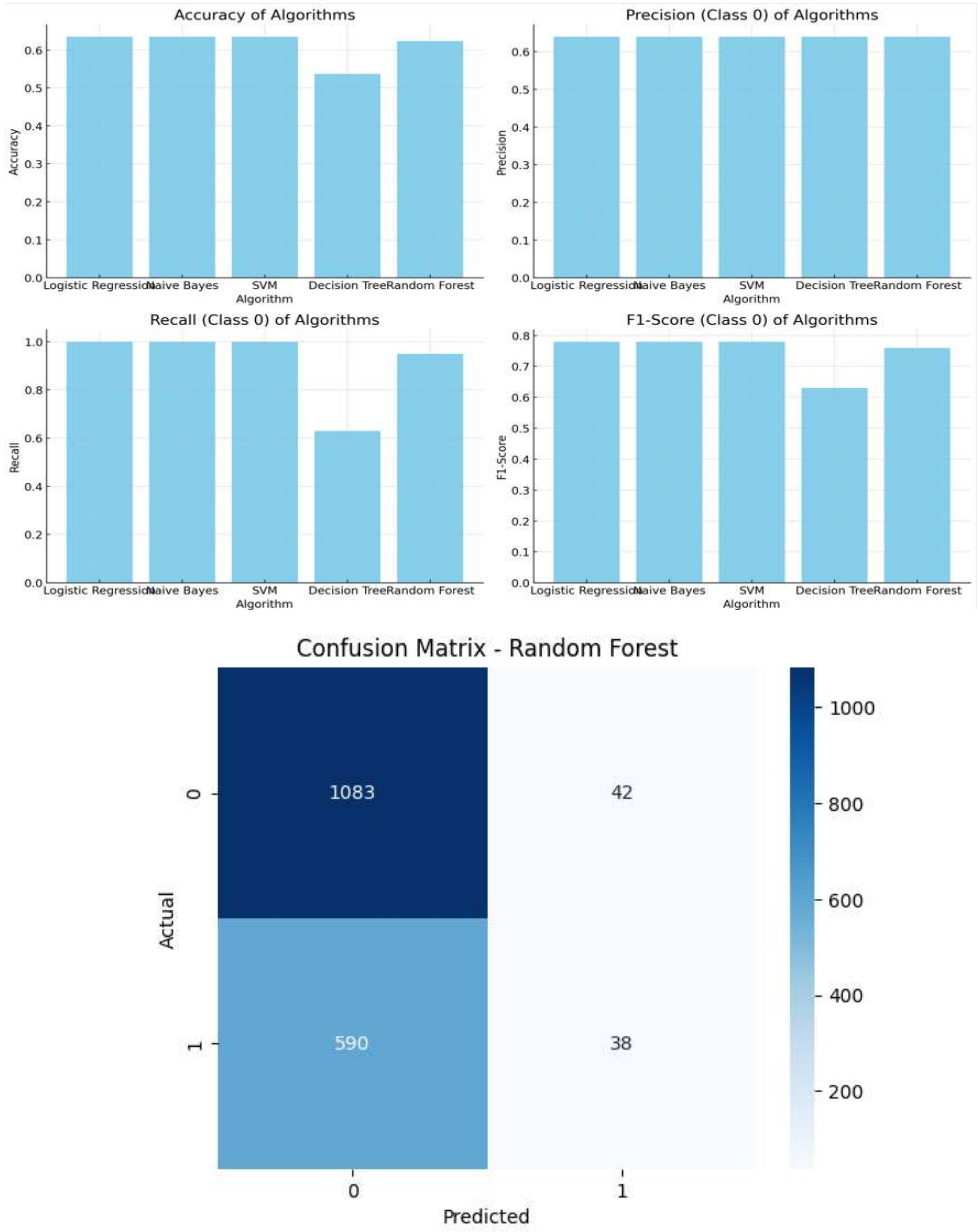
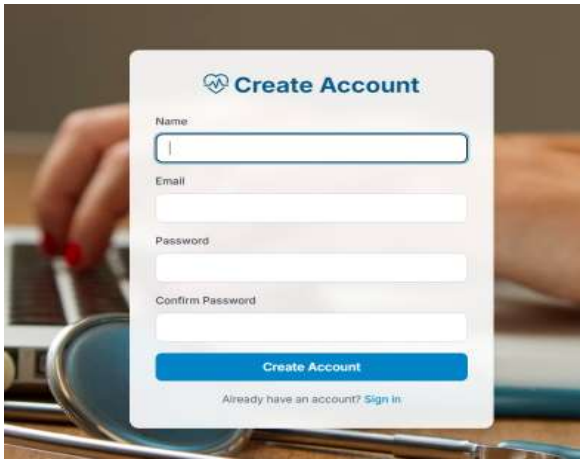
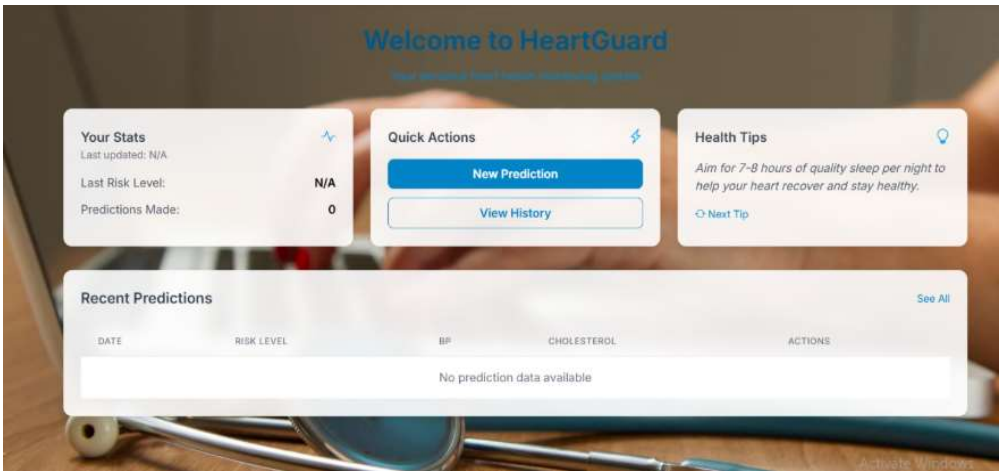


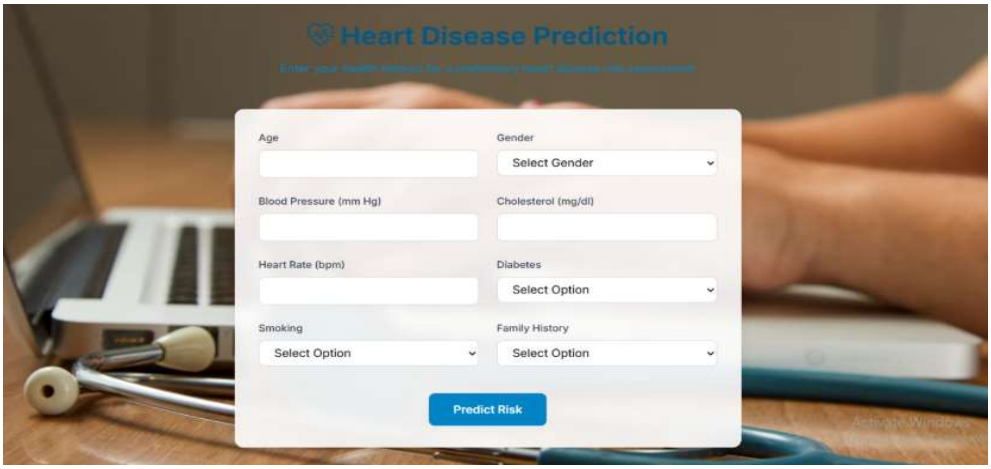
Figure 2: Confusion Matrix for Random Forest Model - Heart Disease Prediction

The confusion matrix displayed above illustrates the performance of a Random Forest model in predicting heart disease, comparing actual values against predicted ones. The matrix shows that the model correctly identified 1,083 instances as negative (True Negatives) and 38 instances as positive (True Positives). However, the model made 42 false positive predictions, where it incorrectly classified negative cases as positive, and 590 false negatives, where it misclassified positive cases as negative. This suggests that the model is more likely to misclassify positive cases, as seen in the relatively high number of false negatives. Despite these errors, the model performs reasonably well, especially in identifying negative cases. Improvements could be made in reducing false negatives to enhance its accuracy in predicting heart disease.

Web application Results:







8. Future Scope

1. **Real-Time Prediction and Monitoring**: The future of heart disease diagnosis will likely involve real-time prediction systems integrated with wearable health devices (e.g., smartwatches, ECG monitors). Continuous monitoring of patient vitals can trigger immediate alerts and interventions.
2. **Integration with Electronic Health Records (EHRs)**: By integrating predictive models with EHR systems, healthcare providers could seamlessly access automated diagnosis tools and personalized treatment plans, improving clinical decision-making.
3. **Hybrid and Ensemble Models**: Future work could focus on combining multiple models (e.g., SVM, Random Forest, and ANN) into a hybrid system that leverages the strengths of each algorithm. Ensemble learning techniques like boosting and bagging could further enhance prediction accuracy.
4. **Explainable AI in Healthcare**: Moving forward, incorporating explain ability frameworks such as LIME or SHAP with complex models like ANN and CNN can improve model transparency, making it easier for healthcare professionals to trust and interpret predictions.
5. **Cross-Domain Application**: While this study focuses on heart disease, the methodologies developed can be adapted to predict other chronic diseases, such as diabetes or stroke, by using similar machine learning techniques and health data, thus broadening the impact of the research.
6. **Deployment in Low-Resource Settings**: Efforts to optimize models for faster execution on lower-end devices or cloud-based solutions can make this technology more accessible to healthcare systems in developing countries, improving global health outcomes.

9. Conclusion

- In conclusion, the heart disease prediction model developed using various machine learning algorithms, including Logistic Regression, Random Forest, Neural Networks, and others, offers valuable insights into the early detection and risk assessment of heart disease. By systematically collecting and preprocessing comprehensive clinical and demographic data, the model can effectively analyze various risk factors and provide accurate predictions regarding the presence or absence of heart disease.
- The developed heart disease prediction model has significant implications for healthcare practice. It enables early detection, risk stratification, and personalized patient care by providing accurate predictions regarding the presence or absence of heart disease. This model empowers healthcare professionals to make informed decisions, develop tailored treatment plans, and allocate appropriate resources based on individual patient profiles. The accurate identification of heart disease helps in reducing morbidity and mortality rates associated with cardiovascular conditions.
- Future research should focus on incorporating more sophisticated feature engineering techniques and exploring other advanced machine learning algorithms to improve prediction accuracy.

10. References

1. D. B., L. M., A. R., J. S. Kallimani, R. Walia, and B. Belete, "A Novel Feature Selection with Hybrid Deep Learning Based Heart Disease Detection and Classification in the e-Healthcare Environment," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Sep. 2022, doi: 10.1155/2022/1167494.
2. M. R. Pulicharla and D. A. Singhal, "Techniques for Machine Learning: Identifying Heart Disease within E-Healthcare through Implementation: Logistic Regression Model".
3. S. Siamala Devi, G. Harini Karthika, and M. Deepika, "Machine Learning based Classification for Heart Disease Identification," *J. Phys. Conf. Ser.*, vol. 1916, no. 1, p. 012174, May 2021, doi: 10.1088/1742-6596/1916/1/012174.
4. A. J. P. L *et al.*, "Medical information retrieval systems for e-Health care records using fuzzy based machine learning model," *Microprocess. Microsyst.*, p. 103344, Oct. 2020, doi: 10.1016/j.micpro.2020.103344.
5. S. A. Alzakari *et al.*, "Enhanced heart disease prediction in remote healthcare monitoring using IoT-enabled cloud-based XGBoost and Bi-LSTM," *Alex. Eng. J.*, vol. 105, pp. 280–291, Oct. 2024, doi: 10.1016/j.aej.2024.06.036.