**Description of activities undertaken**

The diagnosis of heart disease is a complex process that requires analyzing both clinical and pathological data. This method is not only time-consuming but also costly, limiting access to quality medical care for many individuals. According to the British Heart Foundation, heart disease is responsible for a quarter of all deaths in the UK. On average, one person dies from heart disease every three minutes, amounting to approximately 480 deaths per day. Additionally, around 8 million people are currently living with heart disease in the UK. These alarming statistics have driven increased interest among researchers and healthcare professionals to develop accurate and cost-effective methods for predicting heart disease, with the goal of reducing the mortality rate associated with this condition.

As part of my advanced practice, my group and I developed a heart disease prediction model designed to assist medical professionals in determining the likelihood of heart disease in patients. The dataset used in this study was sourced from Kaggle and originates from 1988. It combines data from four different databases: Cleveland, Hungary, Switzerland, and Long Beach V. While the original dataset included 76 attributes, we focused on a subset of 14 attributes deemed most relevant for our analysis. The dataset’s key feature, the "target" field, indicates the presence or absence of heart disease in patients. This is encoded as an integer: a value of 0 signifies no disease, while 1 indicates the presence of disease.

Our work involved several critical steps, including data cleaning, visualization, and preprocessing. After these, we trained the dataset for prediction, by employing various machine learning models, such as Support Vector Machine (SVM), XGBoost, Random Forest, and Stacking. Additionally, hyperparameter tuning was performed to enhance the performance of the models, although the improvement observed was marginal. Following the training process, the performance of each model was evaluated using standard metrics to identify the most accurate model. Among the models tested, the Support Vector Machine and Stacking models achieved the highest accuracy.

**Literature Review**

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors** | **Approach** | **Best Accuracy** | **Dataset** |
| Ouf and ElSeddawy, 2021 [**21**] | Repeated random with random forest | 89.01%(random forest classifier) | UCI cardiovascular dataset (303 patients, 14 attributes) |
| Shorewall, 2021 [**5**] | Stacking of KNN, random forest, and SVM outputs with logistic regression as the metaclassifier | 75.1% (stacked model) | Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes) |
| Waigi at el., 2020 [**12**] | Decision tree | 72.77% (decision tree) | Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes) |
| Maiga et al., 2019 [**7**] | -Random forest -Naive Bayes -Logistic regression -KNN | 70% | Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes) |
| Khan and Mondal, 2020 [**22**] | 1. Holdout cross-validation with the neural network for Kaggle dataset  2. Cross-validation method with logistic regression (solver: lbfgs) where k = 30  3. Cross-validation method with linear SVM where k = 10 | 1. 71.82% (neural networks)  2. 72.72%  3. 72.22% | 1. Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)  2. Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)  3. Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes) |
| Arora et al., 2024 | Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, Decision Tree, Random Forest, XGBoost | 86.89% (Random forest)  78.69% (XGBoost) |  |
| Sharma et al., 2020 **[1]** | Gaussian Naive Bayes, Decision tree, SVM, Random forest | 99% (Random forest)  98% (SVM) | UCI Cleveland Heart Disease Dataset (1025 instances. 14 attributes) |

**Proposed Methodology**

Planned using two ensembling methods;

1. **Boosting** - Boosting is an ensemble learning technique that trains different machine learning models sequentially, with each new model attempting to correct the errors of the previous one. The final prediction is obtained by combining the outputs of all the models. Boosting can help mitigate both underfitting and overfitting by creating a strong ensemble model that captures diverse aspects of the data, even if individual models are weak or prone to bias or variance**.** The two types of boosting we planned using is **Adaboost, LightGBM** and **XGBoost**
2. **Bagging** - One popular implementation of bagging is **random forests**, which combines bagging with random feature selection. In this method, your training dataset is split into random samples to train a collection of decision tree models. Each training job randomly chooses a subset of features for splitting at each node of the decision trees.

**Application of ethical approval**

Reflecting on the ethical approval process, our group sourced the dataset for this project from Kaggle **[3].** Recognizing the importance of ethical compliance, I applied for research ethics clearance through the university's online Ethics Review Process (ERM) on behalf of my group. The process, though straightforward, required careful attention to detail to ensure all ethical considerations were addressed. The application was submitted on September 2, 2024, and I felt a sense of accomplishment when it was approved on September 14, 2024. Receiving the confirmation email (**figure 1**) not only validated our efforts but also reinforced the significance of maintaining ethical standards in research.

***Confirmation of research ethics Clearance***

*Dear  CHARLES OLANREWAJU*

*RE: Comparative Analysis of Machine Learning Techniques in Heart Disease Prediction 2024 Sep 23920 OLANREWAJU*

*Your application has been reviewed and I can confirm that this study has received ethics review Clearance and can proceed as soon as you receive this confirmation.*

*Please note that if you need, in the future, to make any amendments to your study details now that it has been approved, that you can make a request for amendments using the ERM form itself. To do this, go into your original application and create a 'Sub-Form' to request amendments. Details about doing this are provided in the ERM Applicant Guide for 2021-22.*

*Good luck with your study*

*Yours sincerely,*

*Dr Michael Knowles*

*on behalf of*

*SCEDT Research Ethics sub-Committee*

**Figure 1:** Ethical Clearance Confirmation Mail

The purpose of the ethical clearance application is to ensure research is conducted responsibly, safeguarding the rights, safety, and well-being of participants, researchers, and stakeholders. It promotes compliance with legal and ethical standards, assesses and mitigates risks, protects data privacy, and upholds research integrity. By addressing potential ethical concerns before a project begins, the process minimizes harm, fosters trust, and ensures transparency and accountability in research practices.

**Hyperparameter Tuning**

Hyperparameters are adjustable settings configured prior to training a machine learning model. Optimizing model performance requires hyperparameter tuning, which entails experimenting with different value combinations to identify the most effective configuration. **[6]**

Adjusting hyperparameters can greatly influence the training duration, the model's ability to converge, and its overall performance.

**Benefits of Hyperparameter Tuning**

1. Defines the model's behavior and characteristics, influencing how it processes and learns from data.

2. Optimizes the model's performance and training efficiency, ensuring better results and resource utilization.

I handled the hyperparameter tuning of our code in order to improve the performance of the models.

**References**

1. Ouf, S.; ElSeddawy, A.I.B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. *J. Southwest Jiaotong Univ.* **2021**, *56*, 220–240. **[21]**
2. Shorewala, V. Early detection of coronary heart disease using ensemble techniques. *Inform. Med. Unlocked* **2021**, *26*, 100655. **[5]**
3. Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. *Eur. J. Mol. Clin. Med.* **2020**, *7*, 1638–1645. **[12]**
4. Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48. **[7]**
5. Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. *Int. J. Comput. Appl.* **2020**, *176*, 46–54. **[22]**
6. <https://explore.skillbuilder.aws/learn/course/19678/play/128163/aws-ml-engineer-associate-23-refine-models>
7. V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India, 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842. **[1]**
8. Arora, D., Sharma, A. and Agrawal, B.K. (2024) ‘Assessing the impact of various machine learning algorithms for heart disease prediction’, Lecture Notes in Networks and Systems, pp. 453–468. doi:10.1007/978-981-97-0700-3\_35. **[2]**
9. <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction> **[3]**