Project Proposal: Cardiovascular Disease Prediction using Machine Learning (Deep Learning, Hybrid Features, Getting Highest Possible Accuracy) on Cleveland Dataset.

Rifath Ahmed - 001187671

November 12, 2023

A Project Proposal for the Degree of (BSc Computer Science)

1 Introduction

Cardiovascular disease is a severe health problem, and people worldwide suffer from heart disease. In 2019, an estimated 17.9 million people died from cardiovascular disease, accounting for 32% of all global deaths. 85% of these deaths were due to heart attacks and strokes (Organization 2021). The crude mortality of cardiovascular disease was 607.64 million cases in 2020, an increase of 29.01% compared to 2010. However, the age-standardized prevalence rate was 7354.05 per 100,000, an increase of 0.73% compared to 2010 (Tsao et al. 2023). Early detection and treatment of heart failure can improve quality of life and prolong survival. Heart disease can be detectable through medical tests, electrocardiograms (ECG), risk assessment tools, Etc.

(Rath et al. 2021) They used an unbalanced MIT-BIH arrhythmia dataset and PTB ECG to evaluate the HD detection performance of subjects. For this purpose, six classification models were selected, including two ML, three DL and one ensemble model.

(Akbilgic et al. 2021) Investigates the predictive accuracy of the electrocardiogram (ECG) solely in predicting the risk of heart failure (HF) using artificial intelligence processing. ECG artificial intelligence deep learning models could predict future heart failure with comparable accuracy using only standard 10-second 12-lead ECG data from 14,613 participants from the Atherosclerosis Risk in Communities (ARIC) study cohort.

(Yu et al. 2022) They presented a real-time multimodal biosignal (ECG and PPG) –based stroke disease and health monitoring system for older people. The suggested approach uses a machine learning algorithm to predict the prognostic symptoms of stroke in older people in real-time by extracting features from the raw ECG and PPG data based on the peak values of the waveforms.

(Martins et al. 2021) Based on clinical data gathered during a medical examination, they have used data mining techniques (DMTs) to estimate a patient's likelihood of having cardiovascular disease. The Industry Standard Process for Data Mining (CRISP-DM) technique was used for this, and five classifiers—DT, Optimised DT, RI, RF, and DL—were used. The WEKA models were primarily developed using the RapidMiner tool.

Machine learning techniques can extract patterns from massive amounts of medical data and perform predictive analysis. Machine learning provides many advantages over traditional medical procedures, including time and cost reductions, which aid diagnosis. Machine learning is essential in medical diagnosis and healthcare (Qayyum et al. 2020).

This project aims to significantly improve the prediction accuracy of cardiovascular disease outbreaks using advanced machine learning algorithms applied to the Cleveland dataset. The expected high-precision model seeks to improve patient outcomes through timely and precise medical interventions, optimise healthcare resource allocation, reduce associated costs, and create a reliable benchmark for future healthcare predictive analytics, ultimately leading to more comprehensive Health strategies that contribute to the goals of personalised medicine and an informed public.

2 Problem Domain

Machine learning for predicting cardiovascular disease (CVD) is an intricate and significant field of study, but it has its own set of difficulties and potential issues. It can be challenging to determine which characteristics—biomarkers, risk factors, and medical history—are most important for predicting CVD. Working with many features can cause the curse of dimensionality to impact model performance. To enhance performance, a promising technique for Principal Component Analysis in Heart Failure (PCHF) selects prominent features (Qadri et al. 2023). The Relief and Least Absolute Shrinkage and Selection Operator (LASSO) techniques are used to choose appropriate features to predict cardiovascular disease using machine learning (Ghosh et al. 2021). A hybrid genetic algorithm (GA) and particle swarm optimisation (PSO) optimised approach based on random forest (RF), called GAPSO-RF, implements multivariate statistical analysis in the first step to select the most significant features used in the initial population (Mohamed G. El-Shafiey, 2022) (El-Shafiey et al. 2022).

Accurate model training depends on the availability of diverse, high-quality datasets. Predictions can be biased due to incomplete and unbalanced data, especially for rare cardiovascular events. MaLCaDD (Machine Learning-based Cardiovascular Disease Diagnosis) proposes an efficient and highly accurate cardiovascular disease prediction framework. Specifically, the framework first addresses data imbalances and missing values using the Synthetic Minority Oversampling Technique (SMOTE) and the Mean Replacement technique (Rahim et al. 2021).

Ensuring models generalise well to diverse patient populations is essential. Data augmentation adds variance to the training set, making the model more robust and improving its generalisation ability to new data. If a model performs well on training data but poorly on fresh, accurate data, the likelihood of overfitting is reduced. A generative adversarial network (GAN) model is utilised to augment the small and imbalanced dataset (Raniya R. Sarra, 2022). A unique method for overfitting a model by selecting duplicating misclassified examples during the leave-one-out cross-validation (CV) procedure has been proposed (Abdulrakeeb M. Al Ssulami, 2023) (Al-Ssulami et al. 2023).

The quality and quantity of data, feature selection and generalisation, may be the key issues in predicting cardiovascular disease. By addressing these critical issues, we can achieve the highest possible accuracy, ensuring the reliability of the prediction. The accuracy of disease prediction models is vital and provides deep insights into cardiovascular disease's etiology and risk factors. In addition, they enable the research and validation of new hypotheses and thus contribute significantly to the further development of precision medicine in cardiology. This precision in prediction is a clinical tool and a cornerstone for continued innovation and knowledge expansion in cardiovascular science.(Sarra et al. 2022) Has come up with the highest accuracy of 99.3% on the Cleveland dataset. This project will aim to achieve the highest possible accuracy on the Cleveland dataset.

Table 1: Table of recent work on Cleveland Dataset

Study	Method	Accuracy
(Sarra et al. 2022)	GAN-Bi-LSTM	99.3%
(Al-Ssulami et al. 2023)	Bagged-DT + Cleveland-DT-474-10	99.2%
(MahaLakshmi & Rout 2023)	Ensemble (KNN, SVM, DT, NB and LR)	98.41%
(El-Shafiey et al. 2022)	GAPSO-RF	95.6%

3 Methodology

Deep learning, ensemble and hybrid techniques are the most used approaches for cardiovascular disease prediction. For example, GAN-1D-CNN (generative adversarial network and one-dimensional convolutional neural network) and GAN-Bi-LSTM (bi-directional long short-term memory) are used by (Sarra et al. 2022), a hybrid deep learning collaboration with RNN (recurrent neural network) and LSTM (long short term memory) by (Goswami, 2022) (Ghosh et al. 2021) and an ensemble-based IPSO (Improved Particle Swarm Optimisation) used by (MahaLakshmi & Rout 2023).

This project will work with the Cleveland dataset, which consists of 303 records and 14 attributes. This dataset contains a relatively small number of samples and needs to be more balanced, which may result in an overfitted model. To address this problem, a generative adversarial network (GAN) will be modelled to generate more data from the collected datasets. After expansion, it will use correlation coefficient feature selection techniques to select highly correlated features with the target variable. After feature selection, it will use SMOTE on the training set to resolve any class imbalance. We will train the processed data using a recurrent neural network (RNN). To conduct the study, it will be deployed on the Windows platform running Python.

4 Evaluation

Evaluating the model's success in predicting cardiovascular disease includes statistical evaluation and cross-validation. It will use confusion matrix-based metrics (accuracy, precision, recall(sensitivity) and F1 score) for statistical evaluation. After that, we will perform k-fold cross-validation to ensure the model is robust and generalises well across the dataset.

References

Akbilgic, O., Butler, L., Karabayir, I., Chang, P. P., Kitzman, D. W., Alonso, A., Chen, L. Y. & Soliman, E. Z. (2021), 'Ecg-ai: electrocardiographic artificial intelligence model for prediction of heart failure', *European Heart Journal-Digital Health* **2**(4), 626–634.

Al-Ssulami, A. M., Alsorori, R. S., Azmi, A. M. & Aboalsamh, H. (2023), 'Improving coronary heart disease prediction through machine learning and an innovative data augmentation technique', *Cognitive Computation* pp. 1–16.

El-Shafiey, M. G., Hagag, A., El-Dahshan, E.-S. A. & Ismail, M. A. (2022), 'A hybrid ga and pso optimized approach for heart-disease prediction based on random forest', *Multimedia Tools and Applications* **81**(13), 18155–18179.

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R. & De Boer, F. (2021), 'Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques', *IEEE Access* **9**, 19304–19326.

MahaLakshmi, N. V. & Rout, R. K. (2023), 'Effective heart disease prediction using improved particle swarm optimization algorithm and ensemble classification technique', *Soft Computing* pp. 1–14.

Martins, B., Ferreira, D., Neto, C., Abelha, A. & Machado, J. (2021), 'Data mining for cardiovascular disease prediction', *Journal of Medical Systems* **45**, 1–8.

Organization, W. H. (2021), 'Cardiovascular diseases (cvds)'.
URL: *https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(CVDs)*

Qadri, A. M., Raza, A., Munir, K. & Almutairi, M. (2023), 'Effective feature engineering technique for heart disease prediction with machine learning', *IEEE Access* .

Qayyum, A., Qadir, J., Bilal, M. & Al-Fuqaha, A. (2020), 'Secure and robust machine learning for healthcare: A survey', *IEEE Reviews in Biomedical Engineering* **14**, 156–180.

Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A. & Muzaffar, A. W. (2021), 'An integrated machine learning framework for effective prediction of cardiovascular diseases', *IEEE Access* **9**, 106575–106588.

Rath, A., Mishra, D., Panda, G. & Satapathy, S. C. (2021), 'Heart disease detection using deep learning methods from imbalanced ecg samples', *Biomedical Signal Processing and Control* **68**, 102820.

Sarra, R. R., Dinar, A. M., Mohammed, M. A., Ghani, M. K. A. & Albahar, M. A. (2022), 'A robust framework for data generative and heart disease prediction based on efficient deep learning models', *Diagnostics* **12**(12), 2899.

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A., Arora, P., Avery, C. L., Baker-Smith, C. M., Beaton, A. Z., Boehme, A. K., Buxton, A. E. et al. (2023), 'Heart disease and stroke statistics—2023 update: a report from the american heart association', *Circulation* **147**(8), e93–e621.

Yu, J., Park, S., Kwon, S.-H., Cho, K.-H. & Lee, H. (2022), 'Ai-based stroke disease prediction system using ecg and ppg bio-signals', *IEEE Access* **10**, 43623–43638.