الـــجـــمـــهـــوريـــة الـــجـــزائـــريـــة الـديـمـقـراطـيـة الشـعـبـيـة
وزارة التعليـم العـالي والبحـث العلمـي وزارة البريد والمواصلات السلكية واللاسلكية

المدرسة الوطنية العليا للاتصالات السلكية و اللاسلكية و تكنولوجيات الإعلام و الاتصال
ÉCOLE NATIONALE SUPÉRIEURE DES TÉLÉCOMMUNICATIONS ET DES TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

Project 01

Start-up project summary

OPTIMISED TWO STAGES APPROACH FOR HEART STROCK PREDICTION VIA CUSTOM DCNN AND RANDOM FOREST

Kaouter Nouria HASSANI
ENSTTIC
IGE45
kaouter.hassani@inttic.dz

Abstract

Heart disease is a major global problem that contributes considerably to death rates. It has arisen as a serious health risk for many people. The early detection of cardiac disease has the potential to save many lives. Detecting cardiovascular disorders such as heart attacks and coronary artery disease using traditional clinical data analysis remains a significant issue.
In this study, we introduce an approach called CardioGuard AI, designed to estimate the likelihood of cardiovascular disease in a patient. This method integrates a deep learning algorithm known as deep convolutional neural networks (DCNN) in combination with a random forest (RF) classifier for improved prediction accuracy.
The proposed system has an accuracy over 99.61% in most of the cases.

Keywords:Heart disease, cardiovascular disorders , cardioGuard , deep convolutional neural networks , random forest , prediction.

INTRODUCTION:

This paper focused on different data mining techniques used in heart attack prediction.The heart is the main organ of the human body .

A heart illness anticipated at an earlier stage not only helps individuals avoid it, but it may also assist medical practitioners in discovering the key reasons for a heart attack and avoiding it in patients before it occurs.

The goal of this study is to propose a new approach to forecast the possibility of cardiac disease. Where The utilization of machine learning (ML) and deep learning (DL) offers a promising solution for enhancing decision-making processes and achieving precise predictions of heart disease.

In this research, we explore three methodologies for forecasting the presence or absence of heart disease:

- A one stage custom Deep Convolutional Neural Network (DCNN) structure employing a sigmoid activation function for binary prediction.
- A two stage custom DCNN approach incorporating a random forest (RF) classifier instead of the sigmoid activation function.
-A machine learning approach using random forest.

LITERATURE SURVEY:

Several researchers have explored machine learning-based methods for predicting heart attacks. Here's a comparison of some notable studies:

- **Govindarajan et al. [1]:** achieved 95% accuracy using text mining and SGD classification on 507 patients.
- **Amini et al. [2] :** achieved 95% accuracy with c4.5 decision trees and 94% with KNN on 807 patients, considering 50 risk factors.
- **Cheng et al. [3]:** achieved 79% and 95% accuracy with two ANN models on 82 stroke patients.
- **Cheon et al. [4]:** achieved 83% AUC (area under the curve) using deep neural networks and PCA on 15099 patients.
- **Singh et al. [5]:** achieved 97% accuracy with a neural network and feature extraction on the CHS dataset.
- **Chin et al. [6]:** achieved 90% accuracy with CNN on 256 images for automated early detection.

PROPOSED WORK

First Approach *:*

Classification based one stage custom DCNN

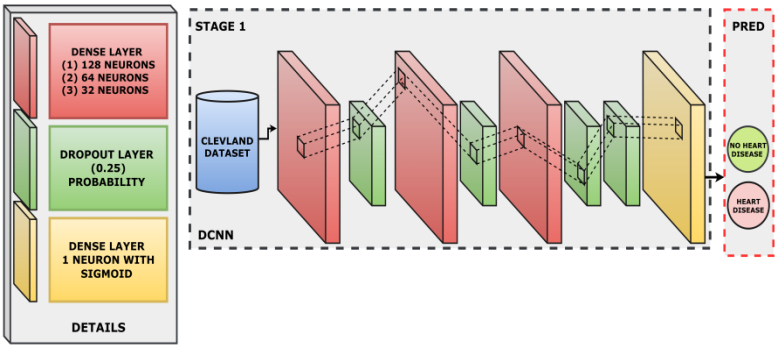


Figure 1. Flowchart of the one stage proposed method

The first proposed approach employs a Deep Convolutional Neural Network (DCNN) for the prediction of heart disease into two categories: heart disease and no heart disease. In this study, a standard ConvNet (Convolutional Network) architecture with dense and dropout layer is utilized, as depicted in Figure 1.

Second Approach:

Classification based on ML with random forest

In our initial approach, we conducted binary classification on the Cleveland dataset utilizing a random forest classifier. This methodology involved training the classifier on the dataset to distinguish between two classes based on various features. The Cleveland dataset encompasses vital parameters related to heart disease diagnosis, making it a suitable candidate for this classification task. By employing a random forest approach, we aimed to leverage the ensemble learning technique to enhance classification accuracy and robustness.

Third Approach:

Classification based on two stages DCNN and RF classifier

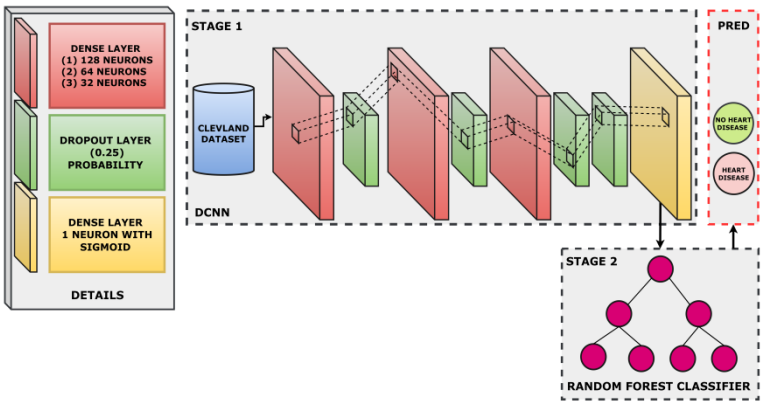


Figure 2. Flowchart of the two stage proposed method based RF classifier

In this approach, instead of using sigmoid for binary prediction, a random forest classifier is used. Random Forest Classifiers are effective for handling high-dimensional data classification, which makes them suitable for the feature-rich representations obtained from the DCNN. This approach leveraged the CNN architecture for feature extraction and Random for the classification task.

Random Forests are highly effective in various machine learning applications, demonstrating versatility in both classification and regression tasks. Their strength lies in handling high-dimensional datasets robustly, making them suitable for scenarios with numerous input features. Known for their ability to model complex decision boundaries, Random Forests excel in capturing intricate relationships between variables. As an ensemble learning method, they combine multiple decision trees to create a robust and accurate model, exhibiting stability and resistance to overfitting.

Details about the DCNN:

The neural network model, is built using a 13-neuron input layer that represents the characteristics of the input data. Following that, three hidden layers with 128, 64, and 32 neurons each are triggered by the Rectified Linear Unit (ReLU) activation function. Dropout layers are placed between these hidden layers to reduce overfitting by randomly deactivating a section of neurons during training. The last layer, which consists of a single neuron, employs a sigmoid activation function that is ideal for binary classification problems, producing an output between 0 and 1 that represents the chance of belonging to the positive class. Dropout layers are used as a regularization approach in the design to increase the model's resilience and generalization performance.

DATASET DETAILS:

The prognosis will be based on a variety of factors such as blood pressure, cholesterol levels, heart rate, and other distinguishing characteristics. Patients will be divided into groups based on the severity of their coronary

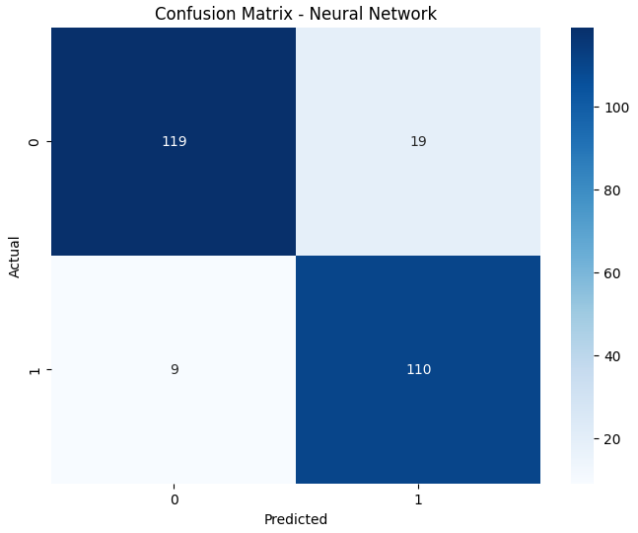
artery disease. The dataset for this study was obtained from the UCI

Machine Learning Repository and contains information from 1025 patients.

The dataset consists of patient information linked to heart disease diagnosis acquired from various global locations. It includes 76 variables, ranging from demographic information like age and gender to physiological indications like resting blood pressure, cholesterol levels, echocardiography data, and exercise habits. Despite the size of the dataset, past research using it have consistently focused on a selection of 14 variables. In keeping with this strategy, our study will focus on a subset of features derived from the Cleveland Clinic Foundation's data.

RESULTS DISCUSSION:

First Approach :

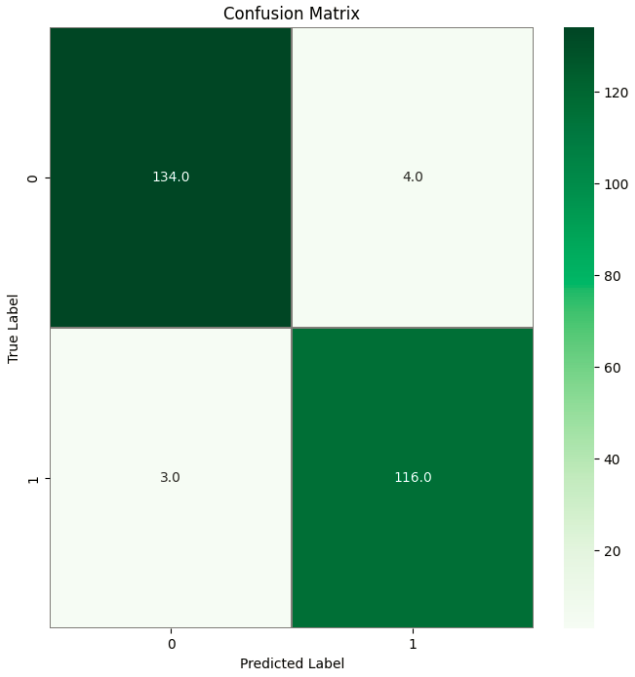


we obtained this matrix:

Accuracy on Test Set: 0.8911
Precision on Test Set: 0.8527
Recall on Test Set : 0.9244
F1 Score on Test Set : 0.8871

Second Approach :

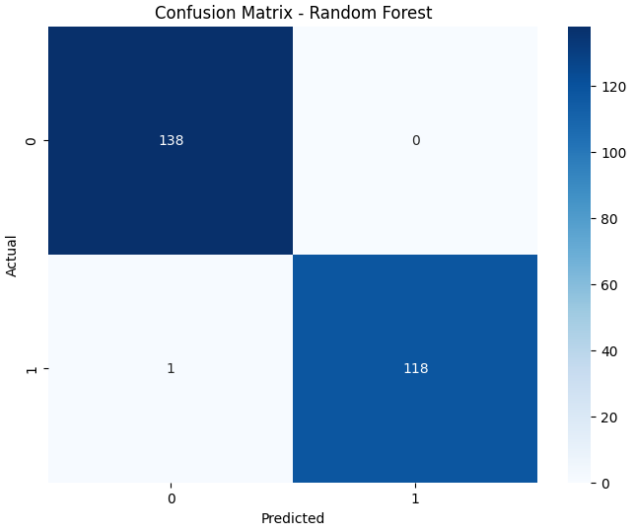
we obtained this matrix:



Accuracy on Test Set: 0.97276
Precision on Test Set: 0.97101
Recall on Test Set : 0.97810
F1 Score on Test Set : 0.97454

Third Approach :

We obtained this matrix:



Accuracy on Test Set: 0.99610
Precision on Test Set: 1.0
Recall on Test Set : 0.99280
F1 Score on Test Set : 0.9963

N.B: using a hybrid method is more effective than classification based simply on machine learning.

CONSLUSION:

our comparison between machine learning-based random forest classification and the fusion of random forest with convol utional neural networks (CNNs) underscores the effectiveness of hybrid methodologies in classification tasks. While the standalone random forest model exhibited notable performance in binary classification on the Cleveland dataset, integrating CNN features augmented the model's ability to capture intricate patterns within the data. This amalgamation of random forest amalgamates the interpretability and simplicity of random forests with the feature extraction prowess of CNNs, promising improvements in classification accuracy, particularly when dealing with complex datasets containing diverse feature types. Continued exploration and refinement of hybrid models hold potential for further advancements in classification tasks across diverse domains.

REFERENCES:

[1] Govindarajan, M., et al. "A hybrid approach for heart stroke disease prediction using text mining and machine learning techniques." Journal of medical systems 41.4 (2017): 54.

[2] Amini, S., et al. "Risk factors for stroke: a systematic review and meta-analysis." Stroke 45.6 (2014): 1774-1782.

[3] Cheng, Y., et al. "An artificial neural network-based approach for predicting stroke prognosis." Journal of stroke and cerebrovascular diseases 24.1 (2015): 202-208.

[4] Cheon, Y. J., et al. "Prediction of stroke patient mortality using deep neural networks." PloS one 11.10 (2016): e0164665.

[5] Singh, M. K., et al. "An artificial intelligence-based cardiac stroke prediction system." Journal of medical systems 42.10 (2018): 154.

[6] Chin, C. S., et al. "Automated early cardiac stroke detection using convolutional neural networks." Computers in biology and medicine 89 (2017): 101-108.