Identifying Heart Diseases using Deep Learning

Problem Statement

About 180 million people die from heart disease, often known as cardiovascular disease, or CVD, each year [1]. It is one of the deadliest and most common causes of death. Given that cardiovascular diseases (CVD) are the most unexpected of all chronic illnesses, determining an individual's unique risk for CVD is crucial to medically preventing early mortality. Early illness detection and precise diagnosis will enhance the patient's outcome [1].

Despite significant advancements in technology and research, several types of stress tests are still used in cardiology. In order to postpone the diagnosis of illness, doctors and other healthcare providers rely on test findings, prior medical history, and other diagnostic [1]. Indications are insufficient to accurately identify a danger and its magnitude. Heart disease may go undetected in such a situation, which would be problematic if it went untreated. Numerous studies have indicated that early illness identification can greatly increase a patient's chances of survival; nevertheless, the methods used today are ineffective in preventing the early diagnosis of diseases that could result in fatalities or abrupt heart attacks [1].

In order to provide experts with a more realistic picture of the risk profile of patients with cardio-vascular diseases, it is imperative that only reliable prediction models be used. It might be challenging to identify a single factor that influences heart health and risk since there are so many different ones [1]. Examples of factors that may indicate the possibility of cardiac issues include blood sugar levels, cholesterol, age, gender, and heart rate. The most difficult instances to identify are those with a group of symptoms rather than just one [1].

This is the application of deep learning. Owing to the multiplicity of variables influencing heart health, a deeper and more intricate comprehension of the variables that operate together to influence the heart is required. A more complicated model should be used since it is very difficult to identify the exact cause of an illness due to the complex nature of the human body. In order to do this, we must comprehend the relationship between the various elements and how they affect the illness. Neural networks, which allow for the understanding of each factor's effect, can be included to achieve this. A deep learning network with several hidden layers may be used to identify and provide answers to a number of queries, like what factors together create a certain kind of heart disease? What is the relative importance of each element affecting the health of the heart? Early illness detection is possible.

This project's primary objective is to use deep learning techniques (such as feature extraction and integration) to enhance the existing illness prediction model. The goal of this initiative is to support the medical and healthcare sectors in early illness detection and risk assessment. A dataset with a variety of characteristics is taken [1].

Motivation

The goal of this initiative is to identify cardiac disease at an early stage. Since cardiac problems are often discovered later in life, using deep learning techniques can identify them early even in the absence of symptoms. Additionally, it can increase the diagnostic' precision and identify the precise illness the patient has. Because the components and their weights are taken into consideration, this model may examine intricate patterns in the data.

Because less intricate procedures are required, diagnosis costs can be lowered by increasing accuracy and identifying the illness early. Tragedy can be avoided by exercising caution beforehand. This has a broad worldwide impact since it may be utilized anywhere. In addition to predicting and detecting the disease, deep learning approaches may also provide fresh insights from the data. It is possible to identify new patterns that will improve diagnostics.

The primary goal of this initiative is to accurately and early identify cardiac problems. Since late detection is the primary cause of CVDs, early detection will lead to early diagnosis and the potential to save countless lives. In order to cure illnesses, this can aid in early detection and give specialists pertinent data.

Literature Review

One of the most significant issues in the world that has to be resolved quickly and accurately is the prediction of heart disease. Numerous research projects were undertaken in an effort to use artificial intelligence to address this issue. The majority of them are created with machine learning methods and other statistical models, which may not have very good illness prediction accuracy. This is because the elements impacting the illness have a deep correlation with one another.

Shah et al. suggested a machine learning model to predict cardiac disorders using certain supervised machine learning techniques. This study's primary goal was to determine whether the patient had any possibility of foreseeing the illness. The broad binary categorization indicated the patient's likelihood of contracting the illness. A dataset including 303 items was utilized for this investigation. There are seventy-six properties per entry. The dataset came from the UCI repository for machine learning. Just 13 predictor characteristics and 1 class attribute—out of the 76 total—were taken and preprocessed. Numerous algorithms, including Naïve Bayes, Decision Tree, KNN, and Random Forest, were employed. The analysis found that KNN had the best accuracy, at ninety percent [2].

An investigation was carried out utilizing AOC-CapsNet, a medical picture classification system created to supplant the current prediction methods. Octave convolution and attentiveness, two modules intended to choose low- or high-frequency input, were employed. AOC-CapsNet is employed in the categorization of images. Seven distinct datasets, each consisting of training, testing, and validation data, were used to test the data set, and findings were achieved. Although the volatility in the dataset caused it to not function well, the findings were still respectable. It was not a well-developed model for all scenarios because of the data imbalance [3].

Gao classified photographs using deep learning models in order to cope with the unbalanced data. Deep learning employs data with labeled classes, and this domain has a large number of datasets with unbalanced data. "Image complexity-based one-class classification," a deep learning method for one-class classification, was developed to stop this. This method makes use of a neural network that was tested and trained using medical imaging data. It was used to four distinct datasets: Hep-2, FFDM, SOKL, and MRI. With 44.4%, 81.8%, 82.3%, and 67.8% for Hep-2, SOKL, FFDM, and MRI, respectively, it has demonstrated excellent outcomes. In terms of AUC values, the results indicated that the MR dataset performed best, with 96.9 for MR, 92.3 for FFDM, 70.3 for SOKL, and 94.1 for Hep-2 [4].

Several classifiers were employed by Fazl-e-Rabbi et al. to forecast cardiac illnesses. The Cleveland dataset consisted of 270 objects with 76 characteristics from the UCI library. Only thirteen of the dataset's properties were used in this study. To predict cardiac diseases, three distinct classifiers were used: support vector machines (SVM), ANNs, and k-nearest neighbors. Using SVM, the classification accuracy was 85.18%. When k was raised to k=10, the accuracy value using KNN increased. By then, its accuracy rate had risen to 80.74%. 73.33% was the ANN's accuracy rate [5].

Manimurugan's primary objective in the study was to use AI and IoMT in conjunction with cardiovascular event prediction. Data is collected via medical sensors attached to patients over the brief interphase. The association approach known as HLDA-MALO—which stands for Hybrid Optimization approach and Join and LDA—was employed to classify the sensor data. The next phase was to classify the echocardiogram pictures using the hybrid R-CNN. The research used annual asset-return data from the Canadian and Brazilian stock markets in addition to the Cleveland dataset from the UCI repository, which is a feature set consisting of 14 characteristics.

We have previously confirmed that the HLDA-MALO models had 96.85% and 98.31% accuracy, respectively, when implemented with normal and abnormal sensor data [6]. The performance measures developed for R-CNN were evaluated; its grades for specificity, recall, precision, F score, and accuracy were 98.06%, 98.95%, 96.32%, 99.02, and 99.15%.Manimurugan's primary objective in the study was to use AI and IoMT in conjunction with cardiovascular event prediction. Data is collected via medical sensors attached to patients over the brief interphase. The association approach known as HLDA-MALO—which stands for Hybrid Optimization approach and Join and LDA—was employed to classify the sensor data.

The next phase was to classify the echocardiogram pictures using the hybrid R-CNN. The research used annual asset-return data from the Canadian and Brazilian stock markets in addition to the Cleveland dataset from the UCI repository, which is a feature set consisting of 14 characteristics. We have previously tested the HLDA-MALO models using normal and aberrant sensor data, and found that they had an accuracy of 96.85% and 98.31%, respectively. The performance measures developed for R-CNN were evaluated; its grades for specificity, recall, precision, F score, and accuracy were 98.06%, 98.95%, 96.32%, 99.02, and 99.15% [7].

In Ravjot's heart disease prediction study, a person's illness status is ascertained by linear regression. This study aims to use logistic regression to estimate the patient's 10-year chance of developing heart disease in the future. The dataset, which includes around 4000 patient records with 15 characteristics, was obtained from Kaggle. a matrix showing the relationships between each factor and other factors. When logistic regression is used, the accuracy comes out to be 85% [8].

Methodology

It is evident from the study and reading of all the articles and initiatives that applying deep learning may be very helpful in predicting cardiac disease. In this experiment, we forecast cardiac illnesses using a deep neural network. To determine if a patient's heart is in excellent or terrible health, we build a neural network.

Dataset

The University of California, Irvine developed Kaggle, which provides the dataset that the suggested technique uses. This is the dataset's outline: Information from 4,000 patients is included in it. There are 76 characteristics total for each patient input, 16 of which are important and required for our model to forecast. The remaining characteristics barely affect heart disease at all. Any null or "nan" values are eliminated if they exist. Next, all the extraneous characteristics are eliminated by pre-processing the data. The model will be more accurate because to the inclusion of more patient data in this data set. The dataset is then randomly shuffled to ensure that there is no trend. An 80:20 split is used to separate the shuffled dataset into training and testing data. The neural network model is trained using the training data, and it is tested using the testing data.

Pre Processing

The practice of eliminating data that is superfluous or not required for the prediction model is known as data preprocessing. By doing this, you may assist the model become more accurate by eliminating ambiguity in the data and nan values. During this stage, any data that contains mistakes or duplicate values is also eliminated. Missing values in any row are eliminated.

Splitting data

The pre-processed data must then be divided into training and testing sets. Eighty percent of the whole dataset is used for training, while the remaining twenty percent is used for testing. To split the data, we must import model_selection.train_test_split and utilize sklearn. The neural network model is trained using the training data, and it is tested using the testing data.

Building the neural network

We must construct the model before we can train the neural network. We may use Keras or TensorFlow to construct the model. There are 15 input nodes because there are 15 characteristics that need to be input with the inputs. There are 15 nodes in the input layer, the top layer, for each input. In order to let the model to learn from all the factors, I want to add two hidden layers in addition to the input layer. Since our primary objective is to identify the factors that contribute to heart disease, it is crucial to include several layers. At minimum, two layers are required to enable the model to understand the relationships between each element.

The model's output layer consists of a single node that represents the sole value used to determine whether or not the patient is ill. The patient's heart is healthy if the output is less than the threshold number, but if it is above, the patient is at danger. Thus, there will be one input layer and fifteen nodes in the neural network. The number of nodes may be decided later, and there will be two hidden levels. One node in the output layer represents the value used to determine whether or not the patient has the illness. To ascertain the risk, we may then incorporate a binary classifier into the project.

We take two numbers, 0 and 1, and put them together. 0 means there is no danger, while 1 indicates there is. I intend to utilize the "Relu" activation function for the activation function and the importable Adam optimizer for the optimizer. Convolution neural networks (CNNs) are what I hope to use as the neural network model, if that's feasible.

Whether or whether the person has an illness will be revealed by the end outcome. There are just two outcomes that can occur: risk or no risk. We can calculate the accuracy, precision, recall, and F1 score using the values. We are able to compute the loss function as well.

Results

The final result is that I used a binary classifier for further analysis and to increase the prediction and accuracy. The accuracy of the model is around 85%.

```
23/23 [==============================] - 0s 2ms/step
Results for Binary Model
0.8442622950819673
              precision    recall  f1-score   support

           0       0.85      0.99      0.91       619
           1       0.45      0.04      0.08       113

    accuracy                           0.84       732
   macro avg       0.65      0.52      0.50       732
weighted avg       0.79      0.84      0.79       732
```

Other values are also indicated in the above picture.

Timeline

15th March 2024 – Project Progress submission

Developing Stage

15th March 2024 - Project Progress

21st March 2024 – Splitting the data into training and testing the data

22nd March 2024 – Building a Neural Network with the above specifications

5th April 2024 – Fitting the model in the current data and testing accuracy and other factors.

Documentation Stage and Submission

19th April 2024 – Writing reports and documentation.

22th April 2024 – Reports and analysis.

27th April 2024 – Project Completion with all the documentation Code and Demo.

Submission

10th May 2024 – Project Submission with all the source code files, dataset and a demonstration video.

References and Citations

1. A Deep Convolutional Neural Network for the Early Detection of Heart Disease. (2022,November 22).

2. Shah D., Patel S., Bharti S.K. Heart Disease Prediction using Machine Learning Techniques. SN Comput. Sci. (2020)

3. Shi G., Wang J., Qiang Y., Yang X., Zhao J., Hao R., Yang W., Du Q., Kazihise N.G.F. Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification. Comput. Methods Programs Biomed. (2020)

4. Ullah A., Anwar S.M., Bilal M., Mehmood R.M. Classification of arrhythmia by using deep learning with 2-D ECG spectral image representation. Remote Sens. (2020)

5. Du N., Cao Q., Yu L., Liu N., Zhong E., Liu Z., Shen Y., Chen K. FM-ECG: A fine-grained multi-label framework for ECG image classification. Inf. Sci. (2021)

6. Rabbi M.F., Uddin M.P., Ali M.A., Kibria M.F., Afjal M.I., Islam M.S., Nitu A.M. Performance evaluation of data mining classification techniques for heart disease prediction. Am. J. Eng. Res. (2018)

7. Manimurugan S., Almutairi S., Aborokbah M.M., Narmatha C., Ganesan S., Chilamkurti N., Alzaheb R.A., Almoamari H. Two-Stage Classification Model for the Prediction of Heart Disease Using IoMT and Artificial Intelligence.(2022)

8. (2023). Ravjot03/heart-disease-prediction. GitHub. https://github.com/Ravjot03/Heart-Disease-Prediction/tree/master