CECS 550 Pattern Recognition

Deep Learning based Conformal Prediction for Cardiovascular Disease (CVD)

Group 04 - Students
Victor Castaneda
Kate Nguyen
Shubham Gupta
Rahul Vishwakarma
Joe Wang

Supervisor
Dr. Mahshid Fardadi

SPRING 2022

Contents

1 Introduction

Cardiovascular Disease describes a variety of conditions that affect the heat and blood vessels. CVD are the leading cause of death globally. According to the World Health Organization, CVD takes approximately 17.9 million lives a year. If discovered early enough, CVD can be managed and treated. Machine Learning methods have been developed to assist physicians in diagnosing CVD early. There are many factors that can contribute to CVD such as High Blood Pressure, High Cholesterol, Weight, Age, Ethnicity, gender, and more. These factors can be used as features in a Machine Learning model. Machine Learning methods, if deployed properly, could potentially help save lives by being able to accurately detect CVD in a patient.

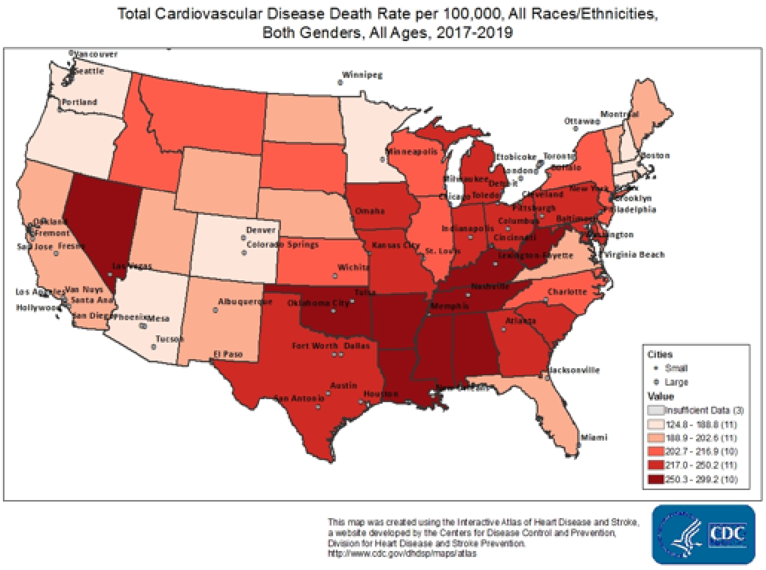

Figure 1: Total cardiovascular disease death per 100,000 from 2017 - 2019

1.1 Motivation

Our goal is to consider the problems Machine Learning faces in the Medical field as well as Prior Machine Learning models. The Models must have high confidence and accuracy in order to be deployed. Once they are deployed they will be able to reliably help save human lives.

1.2 Challenges of machine learning model

Three major challenges Machine Learning methods face in the Medical Field include: Out-of-Distribution Generalization, Incorrect feature attribution, and High-risk decision making. Here we try to discuss each of them for our project.

- Out-of-Distribution Generalization: Machine Learning (ML) methods can be affected by data that is unrelated to the task; Scans of a liver will skew data of a method trained on scans of hearts. Incorrectly prepared data will affect ML performance. Blurry images, incorrect orientation or cropping can affect the processing of data. Methods have been created to allow ML to decide if a data piece should be processed. A number score can be assigned to each piece of data that represents how much it matches the model's training distribution. If the number passes a certain threshold then it is to be processed.

- Incorrect Feature Attribution: Machine Learning models generally take a "path of least resistance" approach in order to select features that correlate with its target output. There is a chance however, that a learning model will select a feature that incorrectly correlates, and it may even cause the model to stop looking for feature in favor of the incorrect one. For Example, if a hospital has many cases in the data set of the target output, the learning model may associate the hospital the data is from with its prediction.

- High-Risk Decision making: Machine Learning methods oftentimes have risk management associated with the decisions the model makes. In the Medical field, the risk is very high. An incorrect diagnosis could cause a patient to receive the wrong or no treatment and could cost the patient their lives. In fact, there are a few real world examples where the ML methods were considered failures.

1.3 Real World Examples - Failures in decision making

There are historical evidence which suggests that machine learning is not a tool to solve all the problem.

- In 2013, IBM and the University of Texas' Cancer Center teamed up to create Watson for Oncology. It was designed with the ambitious goal of curing cancer. In reality, it was a failure that was making incorrect and bad treatment advice. It was later revealed that the machine learning model was trained on hypothetical patients instead of real patients. Whereas physicians are able to use theoretical knowledge to diagnose patients, Machine learning methods must make decisions based on the data it is given.

- Google created a machine learning model designed to detect diabetic retinopathy. The learning model suffered from incorrectly prepared data. The real world data collected in Thailand was too blurry compared to the data it trained on.

2 Existing Approach

Considering all the existing approaches and method used for statistical and deep learning, our prime focus is designing a method which will contribute to Inference and decision making by human beings. One of the limitations of the existing approach is that only accuracy is considered as a metrics for prediction for most algorithms.

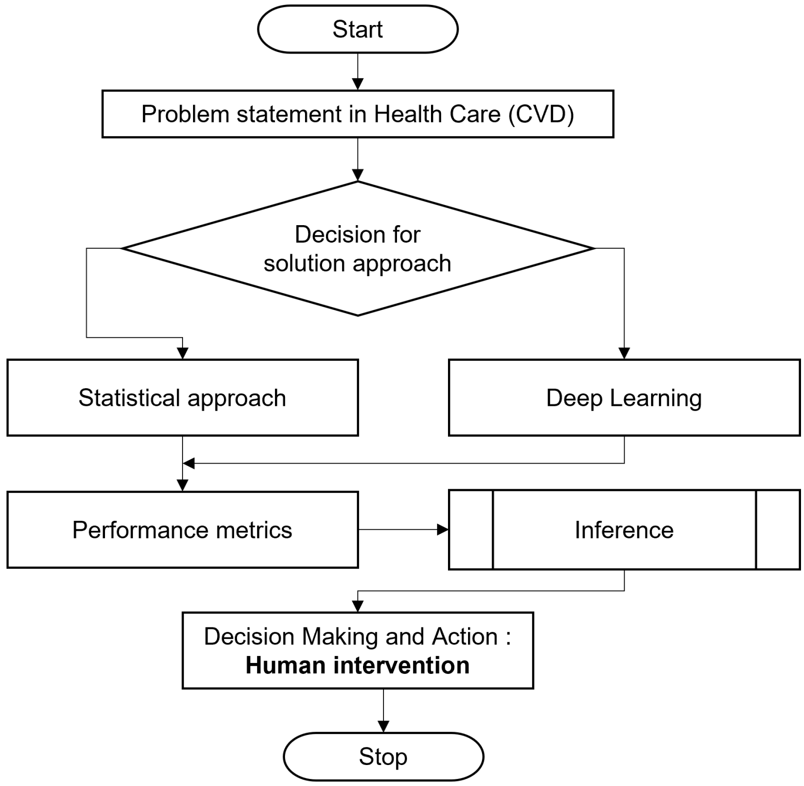

Figure 2: Existing approach to solve machine learning based problem

2.1 Confidence bound

- Statistical Learning Theory allows us to estimate with respect to some confidence level the upper bound on the probability of error. Bounds produced may depend on the VC-dimension of a family of algorithms or other numbers that are difficult to attain for methods used in practice. The bounds usually become informative when the size of the training set is large.

- Traditional statistical methods can be used to compute confidence intervals. Small sample size means the confidence intervals are often too broad to be useful.

- Bayesian methods need strong underlying assumptions

3 Dataset

3.1 Dataset Description

The heart disease dataset from UCI Machine Learning repository, contains 76 attributes, is used in this project. Although, in most of the publications 14 features are used, this model performance will be evaluated by reducing the number of features to 8. The table below describes 8 features of the dataset for Cardiovascular Disease analysis.

Features	Description
ca	number_of_major_vessels
cp	chest_pain_type
thal	thal
sex	sex
oldpeak	ST_depression
thalach	maximum_heart_rate
age	age
chol	cholesterol

Table 1: Heart Disease data description.

3.2 Exploratory Data Analysis

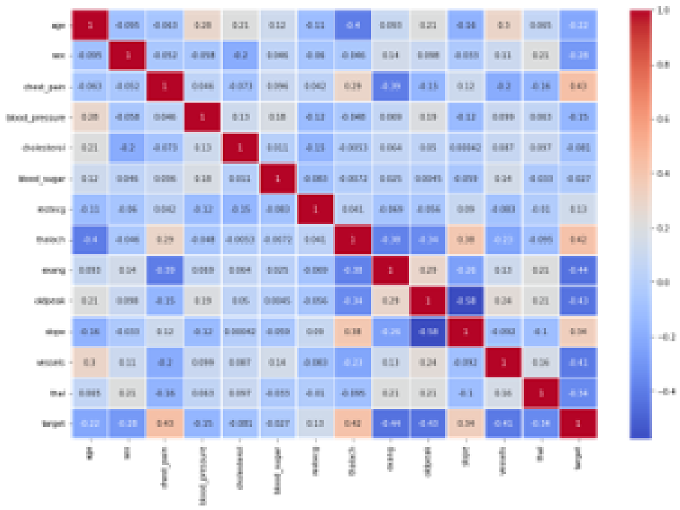

Figure 3: Correlation matrix for the dataset

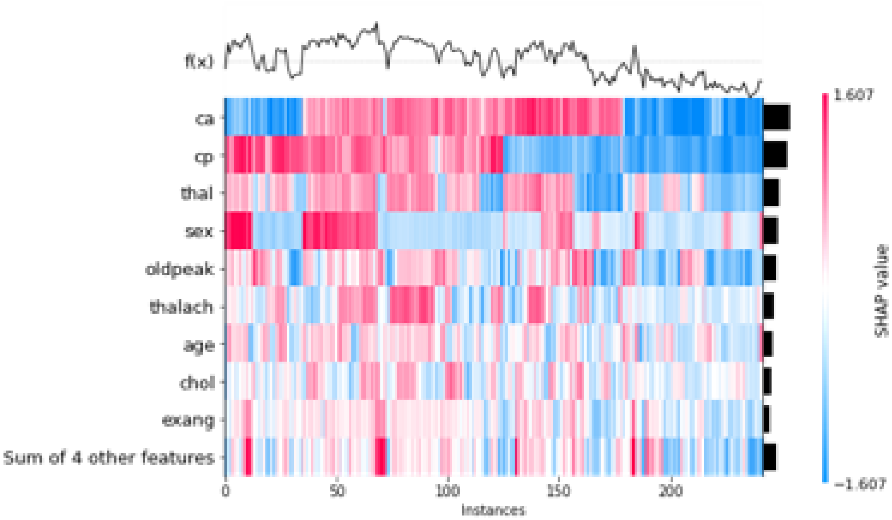

Figure 4: Correlation matrix for the dataset

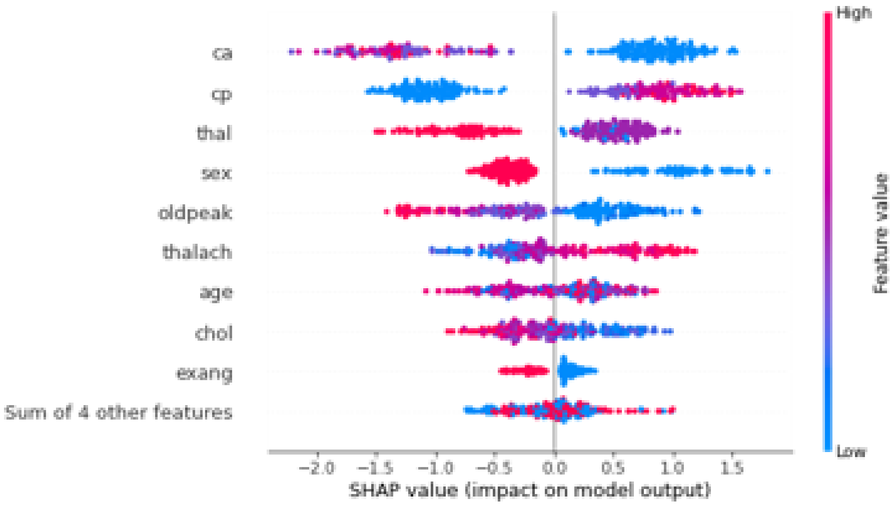

Figure 5: Correlation matrix for the dataset

4 State of the art Algorithms

4.1 Prediction Models

We create predictive models for the heath disease dataset using the below machine learning algorithms

- Logistic Regression
- Random Forest
- K- nearest Neighbor
- Support vector Machine
- Light GBM

- XGBoost

After evaluating the mentioned machine learning models, the accuracy achieved among all models was between 70-85% accuracy. Logistic Regression model reported the best accuracy of 82%. At the same time, Light GBM and XGBoost reported 75.40% accuracy.

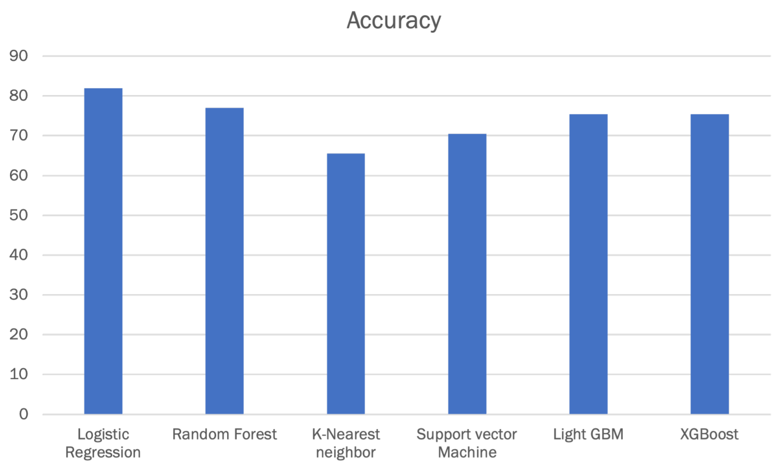

Figure 6: Accuracy of All Models

4.2 Performance Metrics

	Dataset 1	Dataset 2
Logistic Regression	81.96	82.41
Random Forest	77.04	84.61
K-Nearest **neighbor**	65.55	62.63
Support vector Machine	70.49	67.03
Light GBM	75.40	82.41
XGboost	75.40	80.21

Figure 7: Accuracy of All Models

5 Density based Conformal Prediction

We use a density based conformal prediction framework for machine learning classification (Inductive Conformal Prediction), and use density based kernel estimators. First we explain a high level description of conformal prediction and then we show how we can implement density based kernel estimator over a multi layer perceptron (MLP).

5.1 Conformal Prediction

Conformal prediction relies on validation of i.i.d and exchangeability of the dataset. Once it is validated a nonconfrmity score is designed we perform hypothesis testing. Based on hypothesis testing we obtain p-values and consequently confidence and credibility of the prediction.

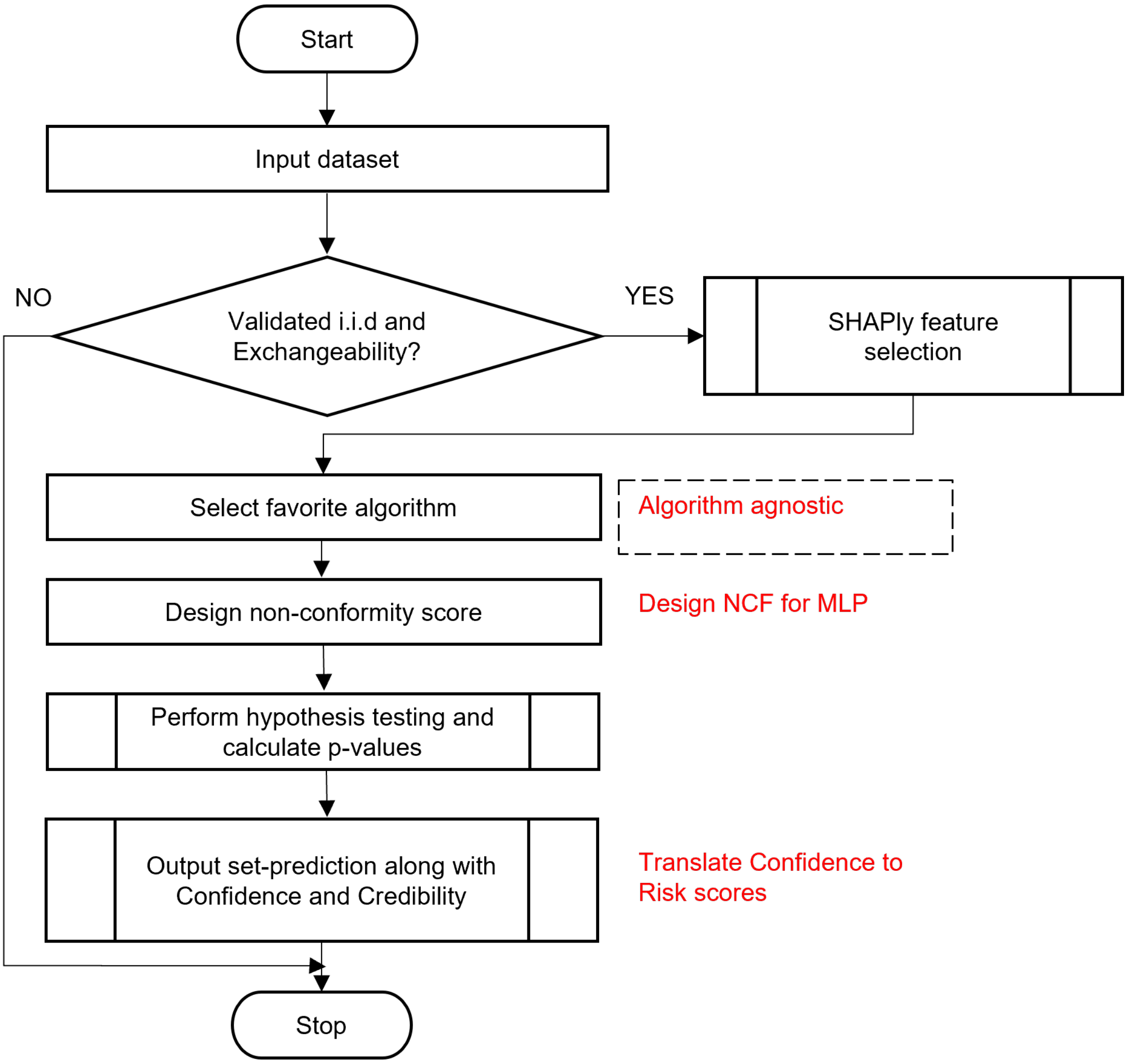

Figure 8: Conformal Prediction Framework

5.2 Proposed Solution

Here we used a gaussian kernel density estimator. The architecture of deep learning models is shown in Fig. 9. It is built following the steps below.

- Use a basic deep learning model depending on the type of data. For Heart disease dataset, this model is a multilayer perceptron (MLP).
- Apply an intermediate dense layer and use it as a feature extractor with a vector of size 50 representing the object, and which will be used later for conformal prediction.
- Add a dense layer to obtain the class predicted by the model (0 or 1).

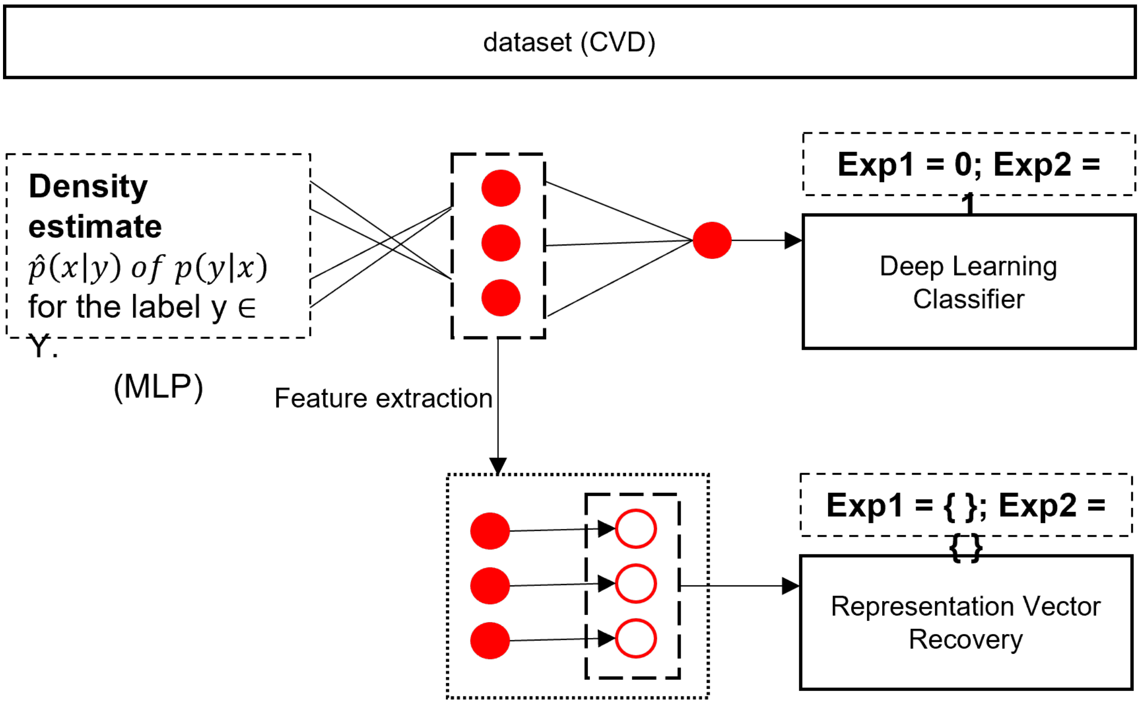

Figure 9: Density based Conformal Prediction Framework for multi layer perceptron

5.3 Experiments

Each data set is divided into proper training, calibration and test sets. A deep learning model dedicated to each type of data is trained on the proper training and calibration sets.
The before last dense layer serves as a feature extractor which produces a fixed size vector for each dataset and representing the object (image, text or vector). These feature vectors are then used for the conformal part to estimate the density.

Result depend on : when is small, the model has high precision and a large number of classes predicted for each observation. On the contrary, when is large, there are no more cases classified as and fewer cases predicted by label.

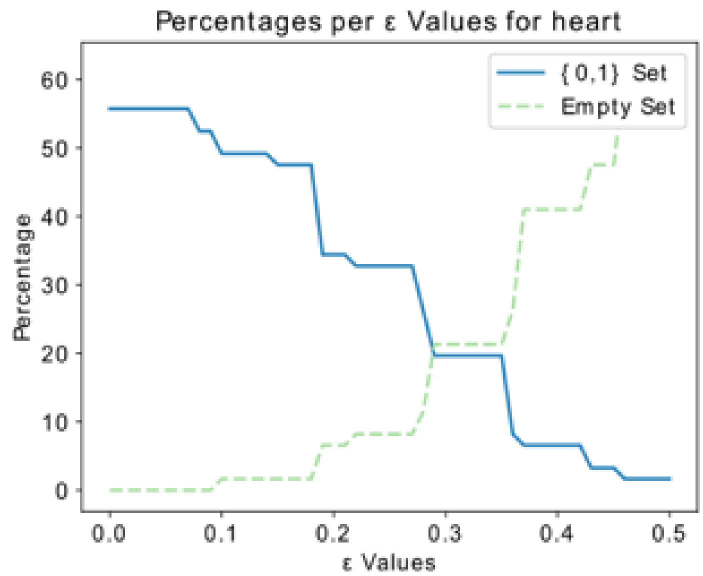

Figure 10: The accuracy and the percentages according to epsilon for Heart

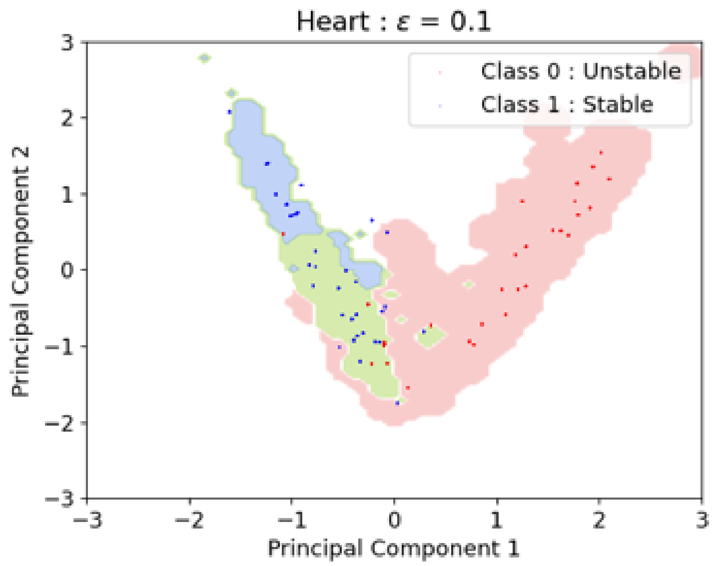

Figure 11: Visualization of the density regions - first two dimensions of a PCA. Non-empty intersection (in green) representing a region of random uncertainty

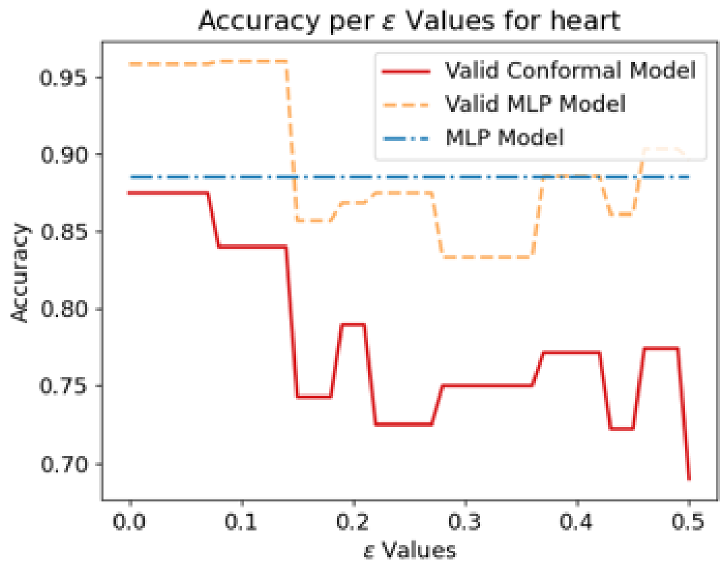

Figure 12: Valid conformal accuracy: the accuracy of the conformal model when one considers only the singleton predictions 0 or 1 (without taking into account the 0,1 and the empty sets).

5.4 Evaluation Metrics

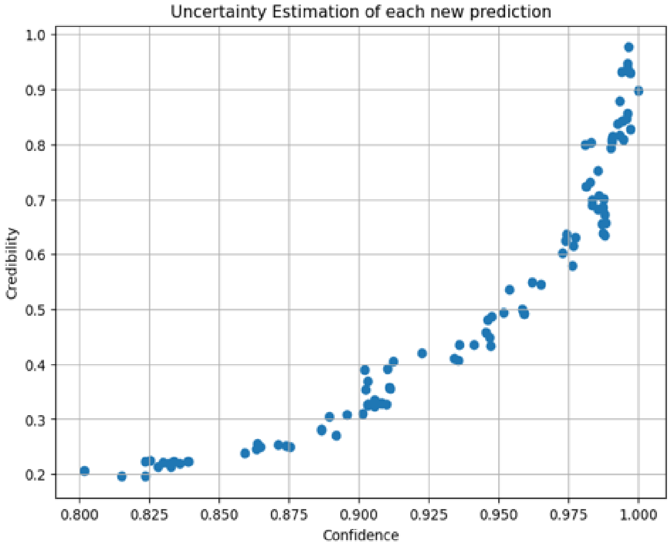

Figure 13: The ideal situation ("clean and easy" data set): max(p0,p1) close to 1; min(p0,p1) close to 0. In this case: both confidence and credibility close to 1

	Label	Confidence	Credibility
0	1.0	0.996830	0.957468
1	1.0	0.906700	0.331908
2	1.0	0.828333	0.223809
3	1.0	0.985969	0.715509
4	1.0	0.985534	0.706089
...	...	...	...
96	0.0	0.996379	0.903441

Figure 14: Low credibility implies either the training set is non-random (biased) or the test object is not representative of the training set.

6 Applications

- Ranking
- Risk sensitive decision making

Extension to Medical image Data

- Cancer prediction
- Depression MRI diagnosis

7 Conclusion

- Trustworthy and reliable predictions
- Algorithm agnostic
- Result is set (rather than point) prediction
- Model can answer – “I don’t know”

8 Appendix

8.1 Source Code

Source code along with figures are hosted on GitHub
https://github.com/rahvis/CECS550.git

8.2 Google Colab

Google Colob source files
https://colab.research.google.com/drive/1JSge4LZxsICUixFMYmvElEI90vXFD4iv?usp=sharing

8.3 Model deployment

The model was deployed locally (localhost) and we recorded the screen for the visualization.
https://youtu.be/yotxB708S_A

9 Acknowledgement

Throughout this project's problem-solving and reporting process, we received a great deal of invaluable insight and guidance from our professor Dr Mahshid Fardadi. We would like to thank her for sharing his expertise with us.

The project, to all of us, is huge and new. This is a tremendous learning experience to tackle a major problem in the medical field and we appreciate very much the opportunity to work on such a project throughout this semester. We thank her support and assistance during the project and reporting process this semester.