LIMITATIONS IN AI-BASED DISEASE PREDICTION SYSTEMS

ARTIFICIAL INTELLIGENCE - CS F407

FINAL ASSIGNMENT REPORT

NAME - ID NO
Jainam Shah - 2019A7PS0096P
Tanmay Parab - 2019A7PS0044P
Aneesh Kabra - 2021A7PS0442P

DATE : 19/04/2023

Overview

The chosen topic is “LIMITATIONS IN AI-BASED DISEASE PREDICTION SYSTEMS.” The recent advancements in technology and Artificial Intelligence, as well as the introduction of AI research in the field of medicine, have shown people the excellent prospects of the usage of AI in healthcare. Deep learning has shown great potential in the field of disease prediction and drug response prediction. With the development in technology, our learning models applied in medicine have also undergone the same progress. This progress has been seen in the continuous improvement in the accuracy of such models in medical disease prediction as well as the overall performance in all aspects.

Via this project, we aim to resolve the issues regarding the scarcity of medical data available for developing high-performance, practically useful prediction models. Medical data, including electronic health records (EHRs), clinical trial data, and other patient-related information, serves as the foundation for training and validating AI algorithms. These algorithms rely on large and diverse datasets to learn patterns, make accurate predictions, and generate meaningful insights. However, the availability of such data for disease prediction is often limited due to several reasons.

Firstly, the privacy and security concerns associated with patient data often restrict the access and sharing of medical data for research purposes. Healthcare providers and organizations are bound by strict regulations which aim to protect patient privacy and ensure data security. These regulations impose limitations on collecting, storing, and sharing patient data, making it challenging for researchers to access comprehensive and diverse datasets for training AI models.

Secondly, the fragmentation and variability of medical data pose a significant challenge to disease prediction using AI. Medical data is often scattered across different healthcare systems, hospitals, clinics, and research institutions, making it difficult to access and integrate into a unified dataset. Moreover, the variability in data formats, coding systems, and data quality standards further complicates the task of aggregating and harmonizing medical data for AI applications. This lack of standardized, interoperable, and comprehensive medical data limits the effectiveness of AI algorithms in disease prediction.

Another challenge is the inherent rarity of certain diseases or conditions. Some diseases, such as rare genetic disorders or certain types of cancers, have low prevalence rates in the general population, making it challenging to collect sufficient data for training AI algorithms. This scarcity of data can result in biased or inaccurate predictions, as AI models may not have enough examples to learn from, leading to suboptimal performance.

Background

Data augmentation is a technique commonly used in machine learning and deep learning to artificially increase the size and diversity of a dataset by applying various transformations or modifications to the original data. Data augmentation aims to create new training examples that are variations of the original data while retaining the same underlying patterns and characteristics. This augmented dataset is then used to train machine learning or deep learning models, improving their ability to generalize and perform well on unseen data.

Data augmentation techniques can be applied to different types of data, such as images, audio, text, and more. Some common data augmentation techniques include:

1) Image augmentation: This can include techniques such as rotation, flipping, scaling, changing brightness or contrast, and adding noise or blur to images.
2) Audio augmentation: This can involve techniques such as changing the pitch, speed, or volume of audio signals, adding background noise, or applying audio effects.
3) Text augmentation: This can involve techniques such as word substitution, deletion, or insertion, changing word order, and generating synonyms or paraphrases.
4) Time-series augmentation: This can include techniques such as time-shifting, resampling, or adding noise to time-series data.

GAN stands for Generative Adversarial Network, a machine learning model used for generating new data samples similar to a given training dataset. GANs consist of two neural networks, a generator, and a discriminator, that are trained together in a process known as adversarial training.

The generator in a GAN takes in random input data, often referred to as noise, and generates new data samples. On the other hand, the discriminator takes in both the generated data from the generator and real data from the training dataset and tries to distinguish between them. The generator and discriminator are trained together in a process where they compete against each other in a game-like manner.

During training, the generator tries to generate data that can fool the discriminator while the discriminator tries to identify whether the data is real or generated correctly. This adversarial process continues iteratively, with the generator improving its ability to generate realistic data samples and the discriminator improving its ability to classify the data correctly. The training process continues until the generator is able to generate data that is indistinguishable from the real data, as determined by the discriminator.

Once trained, GANs can be used for a wide range of applications, including image synthesis, video generation, text generation, and more. GANs have shown remarkable capabilities in generating high-quality, realistic data samples that are difficult to distinguish from real data. They have been used in various fields, including art, design, entertainment, and healthcare.

Literature Review

The papers focus on Artificial Intelligence and its various uses across the medical field. It carefully looks into using AI to help predict diseases in human beings. We look at different machine learning algorithms which make use of extensive data to try and predict the chances of contracting the same disease. We also look at different diseases and the accuracy of different algorithms in predicting them. The papers also shed light on the deep learning model and its applications. We then look at the limitations of the usage of AI and how despite our progress, we have limited accuracy, and using any such model can have catastrophic consequences if any predictions are wrong. In conclusion, the papers then talk about the future possibilities and how we can improve our technology.

The papers look at different machine learning algorithms like -

- Decision Trees
- Logistic Regression
- Support Vector Machines

These papers also detail the concepts behind these algorithms and their various formulae to predict the probability using previous data.

From there, we move to the deep learning model. This algorithm trains multi-hidden layer perceptrons through massive data, extracts useful features, and obtains functions that can effectively represent training data. Deep learning mainly includes Convolutional Neural Networks, Recursive Neural Networks, Generative Adversarial Networks, and many more.

Deep learning models are better than traditional ML models in the following ways-

- High-performance computed processing units (CPUs) and graphics processors;
- Availability of massive amounts of data;
- Development of learning algorithms.

Using Deep Learning, medical images can be classified, segmented, and extracted to assist doctors in completing diagnosis and treatment. This makes the model more comprehensive and stable.

The authors take different models and compare their overall performance in predicting different diseases like diabetes, heart disease, etc. Using deep learning techniques has led to huge improvements in the prediction accuracy from our original systems. It has similar accuracy to non-experienced doctors as well.

The papers then proceed to show the possible limitations of using AI in healthcare -

- Insufficient data to train models
- Not accurate enough for professional use
- Privacy of medical patients and security of their data at risk

One solution for insufficient data is data augmentation, i.e., creating synthetic data based on the limited datasets available.

Problem Statement

Most of the research papers on disease diagnosis and prediction mention the limited availability of medical data (either in the form of images or tabular data) as one of the biggest roadblocks to arriving at better artificial intelligence models that can be put to use in practice in the real world. Thus, to find and propose a solution to this problem, our group was tasked with discovering and testing techniques which could generate new artificial data based on existing limited data.

To approach the aforementioned problem statement, we looked at existing ways to generate data artificially that still resembles the original data. Hence, we found that data augmentation was an extensively used method in AI research in cases where there was loss of data intermittently or only a small dataset existed for rare events. Then we delved into understanding the implementation of data augmentation techniques on image and tabular datasets.

However, such a dataset on some popular disease was needed which had well-formed theories for diagnosis and prediction but had insufficient amount of data so that augmented datasets would show an increase in practicality of the models being built as compared to the original dataset. We found an admissible tabular dataset on Kaggle which has been described in detail later in this report.

Proceeding further, it had to be decided how this tabular data would be augmented. Upon research, the conclusion was made that implementing a GAN-based model (Generative Adversarial Network) would give the best resulting expanded dataset. Hence, the overall assignment covered comparing the performance metrics of different AI prediction models (such as K-Neighbours Classification, SVM, and Random Forest Classifiers) trained on the original and augmented datasets for a particular disease. To make the comparisons easier to understand, the performance metrics were plotted into graphs using python language.

Implementation

The general idea of the assignment was to increase the number of rows in the original dataset or create new datasets entirely from the original dataset. Hence, first the original data file was cleaned and processed before feeding it to a GAN-based model which artificially generated two new datasets. Since the original dataset had a classification variable or a feature for race (namely, Caucasian, Hispanic and Black), it was decided that the entire dataset would be divided into three parts.

Next, the first artificially produced dataset was derived by simply running the GAN model over the three subsets of the original dataset. Hence, it retained a lot of common statistical metrics to the original one but had a greater diversity since the artificial data points generated had a balance of both positives and negatives.

The second artificial dataset was created by feeding only the cases where the person already had some form of heart disease to the GAN model. This time around, the newly generated dataset came out to be less skewed in favor of the majority of people having no heart disease.

Now, the three datasets were input to three different prediction models, namely, a K-Neighbours Classifier, a Support Vector Machine model, and a Random Forest Classifier. All three models were trained on the original dataset for the sake of comparison after testing. As for testing, each of the three models were tested on all three datasets individually creating a total of nine output files with the prediction of heart disease incorporated. Note that the metrics of each model were kept constant for all three datasets for the sake of unbiased comparison.

About the Dataset:

The dataset used for this assignment was picked up from an open-source repository on Kaggle. The prediction variable is the possibility of existence of any number or type of heart disease in a person. This medical dataset was obtained by a survey conducted by CDC in the USA. To elaborate, the dataset comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. As the CDC describes: "Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.". The most recent dataset (as of February 15, 2022) includes data from 2020. It consists of 401,958 rows and 279 columns. The vast majority of columns are questions asked to respondents about their health status, such as "Do you have serious difficulty walking or climbing stairs?" or "Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]". However, it was noticed that a lot of the variables in the dataset were irrelevant for predicting heart diseases. Hence, the original dataset of nearly 300 variables was reduced to just about 20 variables. The variable "HeartDisease" has been treated as a binary ("Yes" - respondent had heart disease; "No" - respondent had no heart disease).

About the Libraries used:

Pandas

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Numpy

NumPy (Numerical Python) is an open source Python library that's used in almost every field of science and engineering. It's the universal standard for working with numerical data in Python, and it's at the core of the scientific Python and PyData ecosystems

Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Create publication quality plots. Make interactive figures that can zoom, pan, update.

Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

SciKit Learn

Scikit-learn is an open source data analysis library, and the gold standard for Machine Learning (ML) in the Python ecosystem. Key concepts and features include: Algorithmic decision-making methods, including: Classification: identifying and categorizing data based on patterns.

Tensorflow

Tensorflow is used for implementing machine learning and deep learning applications. To develop and research fascinating ideas on artificial intelligence, the Google team created TensorFlow.

Keras

Keras is a high-level, deep learning API developed by Google for implementing neural networks. It is written in Python and is used to make the implementation of neural networks easy. It also supports multiple backend neural network computation.

Results and Analysis

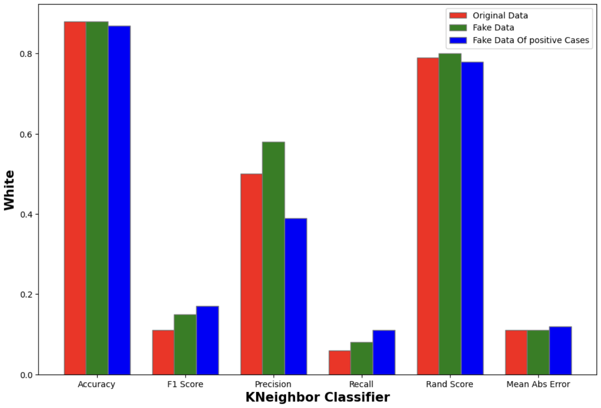
There were **nine** output files created with the final predicted binary variable - “HeartDisease”, obtained by inputting the three datasets (one original and two augmented) to the three pre-trained models (mentioned above). Now, for each of the prediction file data, we calculated the performance metrics of - **Accuracy, F1 Score, Precision, Recall, Rand Score, and Mean Absolute Error**.

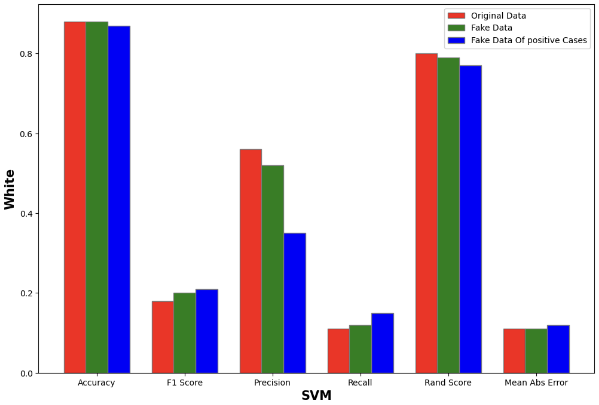
There were no significant differences or trends noted between the Accuracy of the first two datasets but a slight dip was noticed in the third one as the model was less skewed towards negatives.The same could be said about the Rand Score as well. One particular analysis was that the Mean Absolute Error for the third dataset derived solely from positive cases of heart diseases was atleast equal or higher than the original and second datasets. This was expected since the third dataset was supposed to be skewed towards positive cases.

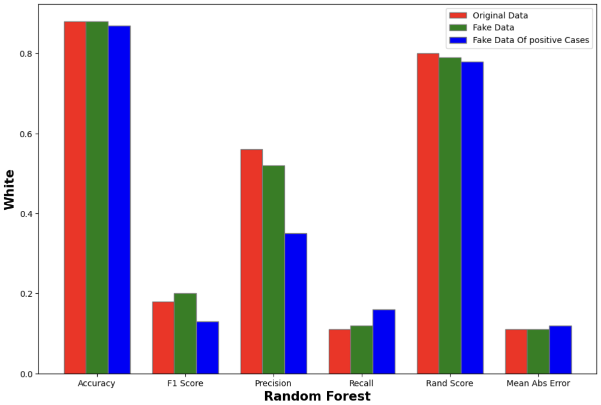
The most important analysis was that despite Precision of the prediction for third and then second dataset was generally lower than the original dataset, the Recall scores were much higher for the augmented datasets which led to a significant jump in the F1 Score of the same. For clarification, the mathematical formulae of some of the performance metrics has been shown below.

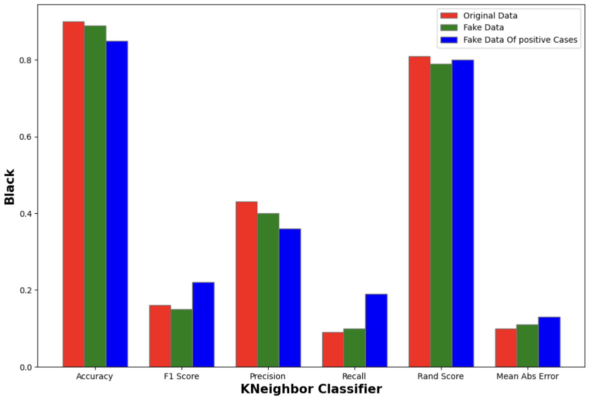
$$\begin{aligned}
\mathit{precision} &= \frac{TP}{TP + FP} \\
\mathit{recall} &= \frac{TP}{TP + FN} \\
F1 &= \frac{2 \times \mathit{precision} \times \mathit{recall}}{\mathit{precision} + \mathit{recall}} \\
\mathit{accuracy} &= \frac{TP + TN}{TP + FN + TN + FP} \\
\mathit{specificity} &= \frac{TN}{TN + FP}
\end{aligned}$$

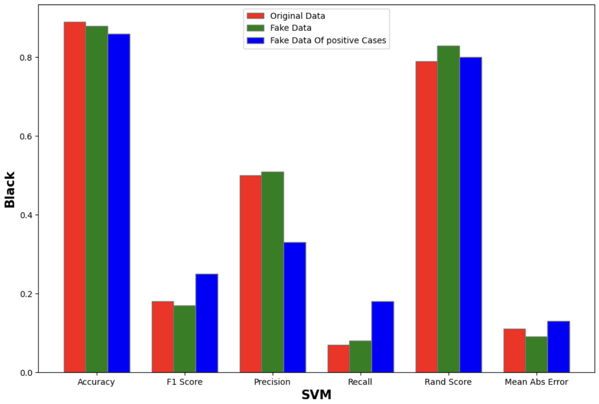
From above, it is clearly evident that F1 Scores of augmented datasets have improved because of a greater increase in Recall as compared to a lesser decrease in the Precision values.

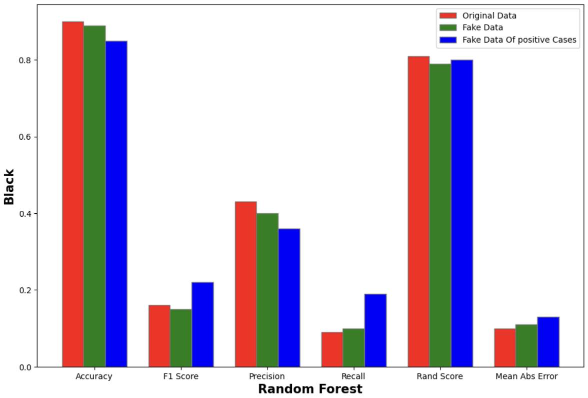

🟥 Original Data
🟩 Fake Data
🟦 Fake Data Of positive Cases
White
0.0
0.2
0.4
0.6
0.8
Accuracy
F1 Score
Precision
Recall
Rand Score
Mean Abs Error
KNeighbor Classifier

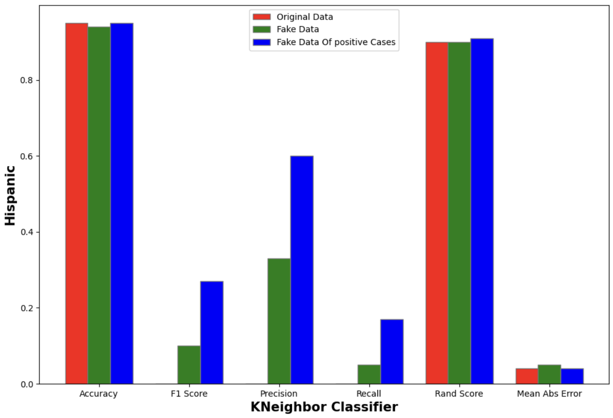

🟥 Original Data
🟩 Fake Data
🟦 Fake Data Of positive Cases
White
0.0
0.2
0.4
0.6
0.8
Accuracy
F1 Score
Precision
Recall
Rand Score
Mean Abs Error
SVM

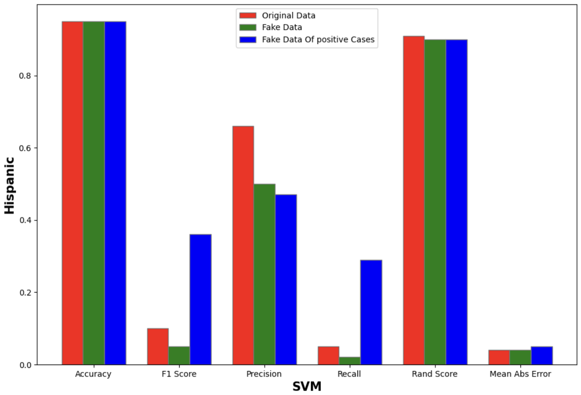

🟥 Original Data
🟩 Fake Data
🟦 Fake Data Of positive Cases
White
0.0
0.2
0.4
0.6
0.8
Accuracy
F1 Score
Precision
Recall
Rand Score
Mean Abs Error
Random Forest


🟥 Original Data
🟩 Fake Data
🟦 Fake Data Of positive Cases
Black
0.0
0.2
0.4
0.6
0.8
Accuracy
F1 Score
Precision
Recall
Rand Score
Mean Abs Error
KNeighbor Classifier


🟥 Original Data
🟩 Fake Data
🟦 Fake Data Of positive Cases
Black
0.0
0.2
0.4
0.6
0.8
Accuracy
F1 Score
Precision
Recall
Rand Score
Mean Abs Error
SVM


🟥 Original Data
🟩 Fake Data
🟦 Fake Data Of positive Cases
Black
0.0
0.2
0.4
0.6
0.8
Accuracy
F1 Score
Precision
Recall
Rand Score
Mean Abs Error
Random Forest


🟥 Original Data
🟩 Fake Data
🟦 Fake Data Of positive Cases
Hispanic
0.0
0.2
0.4
0.6
0.8
Accuracy
F1 Score
Precision
Recall
Rand Score
Mean Abs Error
KNeighbor Classifier


🟥 Original Data
🟩 Fake Data
🟦 Fake Data Of positive Cases
Hispanic
0.0
0.2
0.4
0.6
0.8
Accuracy
F1 Score
Precision
Recall
Rand Score
Mean Abs Error
SVM

Conclusion

In our assignment, we have found a way to augment the original data in such a way that has led to higher Recall scores across our prediction models. This is a very useful function since high recall is an important performance measure in disease prediction models because it measures the ability of the model to correctly identify all individuals with the disease. Recall is the proportion of true positive cases (i.e., individuals with the disease) that are correctly identified by the model among all actual positive cases.

In disease prediction, high recall is particularly important because false negative predictions can have serious consequences. False negatives occur when the model fails to identify individuals who have the disease, leading to missed diagnoses and delayed treatment. This can have a significant impact on patient outcomes, as early detection and treatment are often critical in improving prognosis and reducing morbidity and mortality.

On the other hand, false positive predictions, while still undesirable, are generally less harmful than false negatives. False positives occur when the model identifies individuals as having the disease when they do not, leading to unnecessary medical interventions and costs. While false positives can cause anxiety and distress for patients, they do not have the same level of harm as false negatives.

Therefore, in disease prediction models, high recall is important to ensure that as many true positive cases as possible are correctly identified by the model. This can be particularly important in the early stages of disease when symptoms may be mild or absent, and the disease may be more difficult to detect. High recall can also be important in screening programs or population-based studies where the goal is to identify as many cases as possible to improve overall health outcomes.

In summary, high recall is important in disease prediction models because it measures the ability of the model to correctly identify all individuals with the disease, reducing the risk of missed diagnoses and delayed treatment. While precision and accuracy are also important performance measures, high recall can be particularly critical in disease prediction, where early detection and treatment can have a significant impact on patient outcomes.