AutoHealthX: Automated machine learning and Explainable AI based Cloud functions for Healthcare

Course Name: *Minor Project*

Course Code: *ETMN100*

Abhimanyu Bhowmik

Enrollment No.: A910119819008

B.Tech(Artificial Intelligence)[2019-2023]

Under the Supervision of **Dr. Sukhpal Singh Gill**

Submitted to: Dr Moumita Basu

Amity School of Engineering and Technology

Amity University Kolkata

AutoHealthX: Automated machine learning and Explainable AI based Cloud functions for Healthcare

Course Name: *Minor Project*

Course Code: *ETMN100*

Madhushree Sannigrahi

Enrollment No.: A910119819007

B.Tech(Artificial Intelligence)[2019-2023]

Under the Supervision of **Dr. Sukhpal Singh Gill**

Submitted to: Dr Moumita Basu

Amity School of Engineering and Technology

AMITY UNIVERSITY KOLKATA

Certificate

On the basis of the report submitted by **Madhushree Sannigrahi**, student of B.Tech (AI), I hereby certify that the report **"AutoHealthX: Automated machine learning and Explainable AI based Cloud functions for Healthcare"** which is submitted to the Department of Computer Science, Amity School of Engineering and Technology, Amity University Kolkata in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology(Artificial Intelligence) is an original contribution with existing knowledge and a faithful record of work carried out by her under my guidance and supervision.

To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Sukhpal Singh Gill

Dr. Sukhpal Singh Gill
Assistant Professor
School of Electronic Engineering and Computer Science
Queen Mary University of London, UK

Declaration

I, **Madhushree Sannigrahi**, student of B.Tech (AI) hereby declare that the project titled **"AutoHealthX: Automated machine learning and Explainable AI based Cloud functions for Healthcare"** which is submitted by me to the Department of Computer Science, Amity School of Engineering and Technology, Amity University Kolkata , in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology (Artificial Intelligence), has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition. This project presents the research conducted by the author under the supervision of **Dr. Sukhpal Singh Gill**.

The Author attests that permission has been obtained for the use of any copyrighted material appearing in the report other than brief excerpts requiring only proper acknowledgement in scholarly writing and all such use is acknowledged.

Madhushree Sannigrahi
B.Tech(Artificial Intelligence)[2019-2023]
Amity School of Engineering and Technology
Amity University Kolkata

Certificate

On the basis of the report submitted by **Abhimanyu Bhowmik**, student of B.Tech (AI), I hereby certify that the report **"AutoHealthX: Automated machine learning and Explainable AI based Cloud functions for Healthcare"** which is submitted to the Department of Computer Science, Amity School of Engineering and Technology, Amity University Kolkata in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology(Artificial Intelligence) is an original contribution with existing knowledge and a faithful record of work carried out by him under my guidance and supervision.

To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Sukhpal Singh Gill

Dr. Sukhpal Singh Gill

Assistant Professor

School of Electronic Engineering and Computer Science

Queen Mary University of London, UK

Declaration

I, **Abhimanyu Bhowmik**, student of B.Tech (AI) hereby declare that the project titled **"AutoHealthX: Automated machine learning and Explainable AI based Cloud functions for Healthcare"** which is submitted by me to the Department of Computer Science, Amity School of Engineering and Technology, Amity University Kolkata , in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology (Artificial Intelligence), has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition. This project presents the research conducted by the author under the supervision of **Dr. Sukhpal Singh Gill**.

The Author attests that permission has been obtained for the use of any copyrighted material appearing in the report other than brief excerpts requiring only proper acknowledgement in scholarly writing and all such use is acknowledged.

Abhimanyu Bhowmik
B.Tech(Artificial Intelligence)[2019-2023]
Amity School of Engineering and Technology
Amity University Kolkata

Abstract

Authors: **Abhimanyu Bhowmik, Madhushree Sannigrahi**

Organization: **Amity School of Engineering and Technology**

Title: **AutoHealthX: Automated machine learning and Explainable AI based Cloud functions for Healthcare**

Name of the supervisor: **Dr. Sukhpal Singh Gill**

Month and year of submission: **April 2023**

There is a significant gap between the AI and healthcare industries in the modern world. This is mostly due to a lack of resources and technical expertise. Therefore, our goal in this study is to build a bridge and open up AI to the non-tech population, particularly the healthcare industry. But a significant barrier is that AI needs to be adaptable enough to operate in many contexts with various datasets. To do this, we propose a serverless autoML model with explainable AI. A user-friendly interface for non-professionals has also been created using an interactive web application. Any non-technical user may run the full model, which has a 98 % accuracy rate. AutoHealthX would enable us to utilize the immense potential of AI for enhancing the healthcare sector and providing better service to patients.

Acknowledgements

Words cannot express my gratitude to my mentor and senior scientist Dr. Sukhpal Singh Gill for his invaluable patience and feedback. Without the assistance of Amity University Kolkata, which provided me with the knowledge and experience I needed to conduct the study in this specific area, I would not have been able to go on this adventure. Additionally, this endeavour would not have been possible without the generous support from the Queen Mary University of London, UK who provided me with this research opportunity.

I am also grateful to the professors of Amity University Kolkata, especially Dr Moumita Basu and my classmates from the University, for their help, feedback sessions, and moral support.

Lastly, I would be remiss in not mentioning my family, especially our parents. Their belief in me has kept my spirits and motivation high during the entire process.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
EDA	Exploratory Data Analysis
GCP	Google Cloud Platform
XAI	Explainable AI
AutoML	Automated Machine Learning
SMBO	Sequential Model-based Optimization
LIME	Locally Interpretable Model-Agnostic Explanations
SP-Lime	Submodular Pick Locally Interpretable Model-Agnostic Explanations

Chapter 1

Introduction

The set of practices aiming to base decisions on the analysis of data instead of intuitive insights can be used to define data-driven decision-making. When compared to the conventional ones, businesses that implement data-driven decision-making processes are more financially beneficial and productive [1]. The outcomes of recent AI research projects serve as the foundation for many decision-making tools [2]. The development of machine learning techniques is largely responsible for the success of AI-based tools [3]. This is especially encouraged by the availability of sizable datasets on different real-world features as well as the rise in computational gains, which are typically attributed to the potent GPU cards [4].

One industry produces enormous amounts of data is healthcare using various sensors and health-monitoring devices to collect data. The possibility of developing AI-based processes and models that provide high-value insights for healthcare professionals may be presented by the availability of such data associated with the health of millions of patients [5]. However, creating such procedures and models requires knowledge and skills in data science and artificial intelligence (AI), which aren't always present in hospitals and nursing homes. The work in this paper aims at bridging the gaps between healthcare professionals and AI expertise.

The need to create sophisticated AI models with previously unheard-of performance levels has progressively given way to a rising interest in alternative design elements that would improve the usability of emerging products. Complex AI models lose a part of their practical effectiveness in a wide range of application domains [6]. The main cause is that AI models are frequently created with a performance-focused approach, neglecting other significant—and occasionally crucial—aspects like accountability, transparency, and justice. The AI models are typically "black boxes" since there is no explanation provided for the elements that are projected to perform well; as a result, they simply allow for the prominent display of input and output parameters while hiding the visibility of the intrinsic relationships between those parameters. It is advantageous to have some explanations of individual predictions that are recognised using an AI system, more specifically in an automated environment, because these applications may include crucial decision-making.

The current research aims to develop a transparent and self-explanatory AutoML system that uses serverless computing to propose the best machine-learning configuration for a particular issue and to trace the reasoning behind a recommendation. It could also make it possible to interpretably and credibly examine the prediction outcomes.

1.1 Key Contributions

1. Proposing a system called AutoHealthX for efficient explainable machine learning model building and implementation in the healthcare domain
2. Finding the most accurate and responsive machine learning algorithm for chronic as well as infectious diseases like diabetes, heart disease, breast cancer and COVID-19.
3. Building a prototype web application to provide a visual display for data, models and explainability of results.
4. Implementing the infrastructure on Google Cloud Functions to increase real-time efficiency.
5. Emphasizing prospective future directions.

1.2 Organization of the Report

The remaining portions of the paper are structured as follows: The closely related works regarding ML-based data analytics solutions and the requirement for transparency in order to build confidence in AI models are covered in section 2. The proposed framework is described in section 3 along with how its various elements work together to accomplish the desired outcomes. The results obtained using a few test cases are discussed in section 4. Section 5 demonstrates the AutoHealthX application. Finally, section 6 and 7 concludes the paper's contents and provides an overview of its future directions.

Chapter 2

Literature Review

Numerous AutoML systems have been created in recent years that offer partial or full ML automation, including systems like Auto-sklearn [7], TPOT [8], Auto-WEKA [9], and ATM [10] and commercialised software like Google AutoML[1] , RapidMiner[2] , DarwinAI[3] , and DataRobo[4]. These methods include automatic feature engineering [11] [12], automatic model selection [13] [14], automatic hyperparameter tuning [15], and automatic data preparation [16] [17]. Some methods make an effort to automatically select an ML algorithm while also optimising its hyper-parameters.

Many of the AutoML solutions, some of which are the results of contests from 2015 to 2018, were developed in the last few years. The ChaLearn AutoML Challenges[5] primarily focused on automating the solution of supervised machine learning tasks under certain computing restrictions. These computational restrictions varied significantly across tasks, but they were typically time (about 20 min for training and assessment) and memory consumption restrictions. Guyon et.al [18] reviews a thorough examination of the AutoML difficulties from 2015 to 2018.

In essence, neural architecture may be considered a specific kind of indifferentiable hyperparameter. Hyperparameter optimization is one of these activities that most directly relates to the approach we suggest in this study. Grid search and random search [19] are two of the simplest techniques to find a suitable configuration of hyperparameters from a list of options without considering past results. There are various sets of approaches

[1]https://cloud.google.com/automl
[2]https://rapidminer.com
[3]https://darwinai.com/
[4]https://www.datarobot.com/
[5]http://automl.chalearn.org/

that are often used in hyperparameter optimization. Being one of the most well-known Sequential Model-based Optimization (SMBO) [20] techniques that can take advantage of historical data, Bayesian optimization [21] uses the Gaussian method for prototyping the surrogate function that roughly imitates the relationships between hyperparameters and their desired outputs. All of these techniques are, however, black-box optimised. The single study on AutoML for graph representation [22] employs the Gaussian Process to determine the performance of the hyperparameters, but it scarcely explains how individual hyperparameter affects the performance of the model or why a certain value is picked for a hyperparameter to execute the subsequent assessment trial.

Artificial intelligence (AI) systems that can give human-understandable explanations for their activities and output are referred to as explainable AI (XAI) [6]. By their very nature, end users may be curious as to how and why systems reach to any conclusion [23]. They are seen as "black boxes" when the sophistication of AI algorithms and systems increases [24]. Growing complexity may lead to a lack of openness that makes it difficult to comprehend these systems' logic, which has a detrimental impact on users' faith in them.

Chapter 3

Methodology

In this section, the description of the proposed approach used to achieve the objectives of the work has been discussed, which includes AutoML models, and the usage of Explainable AI to demonstrate the working of the ML models, as well as the server-less architecture of the prototype.

3.1 Dataset

Firstly, the “Breast Cancer Wisconsin (Diagnostic) Data Set” by “UCI ML Repository” is implemented on the novel methodology for the paper [25]. The dataset contains tabular data with 32 features and over 569 data points. A fine needle aspirate (FNA) of a breast lump is used to generate the features from a digital image in 3-dimensional space as described by Bannett et al. [26]. They characterise the properties of all the observable cell nuclei in the image. Every data point is classified into either Benign(B) or Malignant(M) class.

Secondly, the architecture is applied to the “Heart Disease Cleveland dataset” Dataset by “UCI ML Repository” [27]. The dataset constitutes over 300 patients’ data with 75 attributes, However, only 14 of the feature are taken into consideration for determining whether a patient has heart disease or not.

Thirdly, the “Diabetes dataset”, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, is used in this paper [28]. The goal is to determine if a patient has diabetes based on diagnostic parameters. The implemented Diabetes dataset is a subset of an enormous dataset with 10 characteristics and 768 instances. All patients are Pima Indian females who are at least 21 years old.

Lastly, the "Covidgr Dataset" by Tabik et al. [29] is executed in the model. The dataset contains about 26 attributes and 800 samples with various symptoms in a wide range of age groups. The goal of the model is to classify whether the patient is Covid positive or not.

3.2 System Architecture

The experimental architecture was developed using Google Cloud Functions (1st Gen), a serverless solution. The proposed experiment was conducted using cloud functions in the Python 3.9 runtime. Google Cloud Storage is also used for storing the data and its respective output. Cloud Storage events are used as triggers for the respective cloud functions.

In the proposed architecture, two subsequent cloud functions were used: (1) to generate the AutoML model and predict the results; and (2) to explain the predicted results. The open-source Auto-keras python library is used for the first part, the AutoML model, and the lime library is used for the second part, the model explanation. In place of standard Lime explanations, SP-Lime (Submodular Pick Locally Interpretable Model-Agnostic Explanations) is used to explain the model's global decision boundary over a sample set of observations.

3.3 Data Preprocessing

Case-specific and automated data pre-processing was a considerable challenge while implementing the system in a cloud platform. This section discusses the implementation of Pandas Profiling[1] and Feature Tools[2] for automated and rapid EDA and Feature Engineering respectively.

[1] https://pandas-profiling.ydata.ai/docs/master/index.html
[2] https://www.featuretools.com

3.3.1 Automated EDA using Pandas Profiling

Exploratory data analysis(EDA) is a vital stage in constructing any impressive model. EDA involves finding outliers, spotting missing values, figuring out how skewed the datasets are, converting categorical variables, and overall understanding the underlying characteristics and ways to apply them in models.

Pandas Profiling is a user-friendly open-source Python tool for automated exploratory data analysis. It generates a data frame report in a range of different formats. Although the pandas df.describe operation is good, it doesn't give a full data frame report. Pandas profiling was implemented in the system architecture for automated and rapid analysis of data.

3.3.2 Automated Feature Engineering using Feature Tools

Feature engineering is the process of creating and adding new features, or variables, to the dataset in order to enhance the effectiveness and precision of the machine learning model. Case-based knowledge and accessible data sources serve as the foundation for the most efficient feature engineering. Without requiring any human input, automated feature extraction employs deep networks or specialised algorithms to automatically extract characteristics from images or signals.

Featuretools is a free and open-source Python architecture for automated feature engineering. It generates features automatically from relational and temporal information. DFS is utilised for implementing automated feature engineering. For machine learning and predictive modelling, one can construct useful features by combining the raw data with information about the data.

3.4 AutoML

The time-consuming and iterative activities required in developing a machine learning model may be automated using Automated Machine Learning or AutoML. It provides a diverse range of approaches to help those with little background in machine learning access this technology. It attempts to minimize the requirement for experienced individuals to create the ML model. Additionally, it helps to increase productivity and promote machine learning research [5].

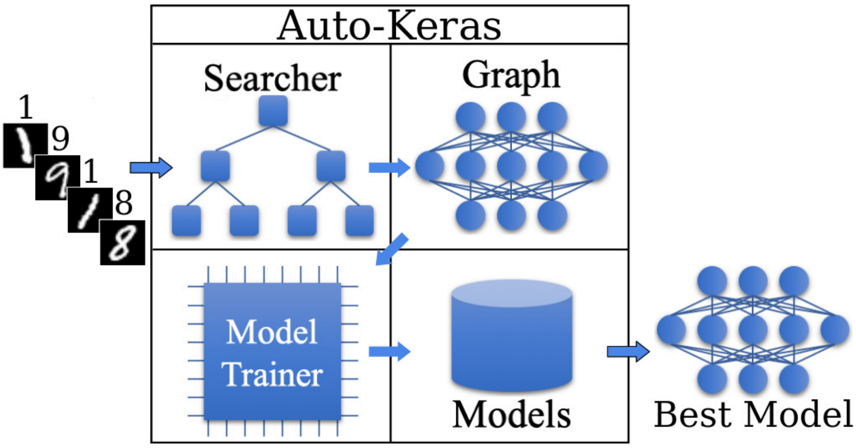

FIGURE 3.1: Auto-Keras Architecture

To properly comprehend automated machine learning, we must first understand the life cycle of a data science or machine learning project. A data science project's lifetime typically includes many stages, including data cleaning, feature engineering, model selection, parameter optimization, and model validation. Even though technology has improved so much, all of these procedures still involve manual labour, which takes time and calls for several data scientists with the necessary skills. For non-ML professionals, it is quite challenging to do these jobs because of their intricacy. The demand for automating these activities has increased because of the rapid development of ML apps, which will make it easier for those without technical expertise to utilise them. Consequently, automated machine learning was developed in order to fully automate the process, from data cleansing through parameter optimization. Not only does it save time, but it also performs fantastically. The overall structure of AutoML is provided in figure 3.1

3.4.1 AutoKeras

A Keras-based AutoML system is called AutoKeras. It is created by Texas A & M University's DATA Lab. Making machine learning accessible to everyone is the aim of AutoKeras. It offers modular building pieces to carry out the architectural searches as well as high-level end-to-end APIs like ImageClassifier or TextClassifier to address machine learning challenges in a few lines.

In this paper, AutoML is considered the first cloud function. It is launched when the clean data is uploaded to the designated bucket. This will create five different models from the training set (which makes up 70% of the total data), train them for 100 iterations, and select the model with the highest accuracy for the given dataset. The performance matrices

(Confusion Matrix, Classification Report) will be generated after the chosen model has been evaluated on test data (30% of the supplied data). The model itself, all training and testing data, performance matrices, and more will be exported into separate cloud buckets. The TensorFlow-Keras standard format for exporting ML models,*.h5*, is used to export the architecture.

3.5 Explainable AI

In the field of machine learning, "explanations" at different levels offer insights into various elements of the model, from knowledge of the learnt representations to the identification of various prediction techniques, general trends and patterns, as well as the evaluation of the general model behaviour [30]. The two types of model explainability are global explainability and local explainability. When a model is globally explainable, users may infer its meaning from its general organisation. Local explainability only takes into account a single input and seeks to understand why the model chooses a certain course of action.

3.5.1 Local Interpretable Model-agnostic Explanations (LIME)

The LIME methodology proposed by Ribeiro et al. [31] generates local explanations of classifier f predictions by fitting a simpler, interpretable explanation model g locally around the data point x to be explained. The overall structure of LIME is given at figure 3.2To maintain interpretability in the generated explanations, LIME represents the data in a way that is comprehensible, locally accurate, and model-neutral. These explanations are simpler because they demonstrate a closer relationship between the input and prediction [32].

For instance, let the grey-scale value vector of pixels in an image be $x \in \mathbb{R}^d$. A comprehensible representation of the initial dataset is used to fit the XAI model. The presence or absence of pixels in the picture might therefore be represented by a binary value vector as an interpretable representation of $x' \in \{0,1\}^{d'}$. (absence refers to having the value of a background colour, e.g., white). As a result of resolving the optimization issue, the LIME explanation $\hat{g}$ is generated.

$$\hat{g} = \underset{g \in G}{argmin} L(f, g, \pi_x) + \Omega(g)$$

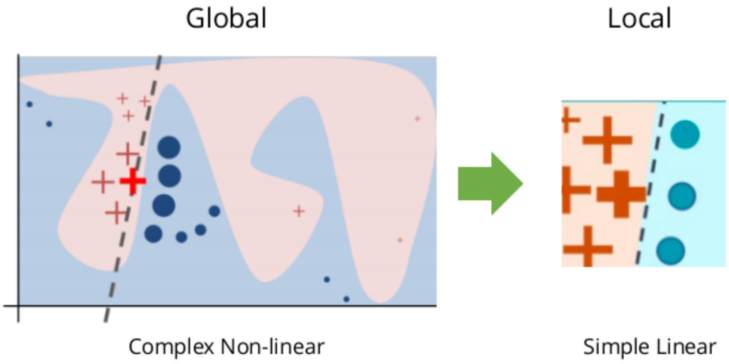

FIGURE 3.2: Structure of LIME.

Here,
G = the explanation model family,
L = loss function,
π_x = the locality around x,
Ω = complexity penalty.

Practically, G is the collection of linear regression models, with Ω limiting the expanse of explanatory features that can possess regression weights other than zero (even if several explanatory models may be employed). The weighted L2 distance is assumed to represent the loss function as:

$$L(f, g, \pi_x) = \sum_i \pi_x(z_i)(f(z_i) - g(z_i'))^2$$

where the sum passes over a collection of selected perturbed points around x, $(z_i, z_i')i = 1, ..., m$, where,
z_i = a disturbing data point in the initial dataset,
z_i' = the corresponding explainable version;

Here, $\pi_x(z_i)$ assigns a weight to each sample according to how similar they are to the point x, which is used to explain the classification result.

The model along with training and test data is received by the second cloud function when they are generated and uploaded to their respective cloud storage. The function generates SP-Lime explanations. For generating the explanations, a random 20 test data samples are chosen, among them, 5 results were generated. The combined graphs were then uploaded to a cloud storage bucket as a single HTML file. Any client-side web or mobile application can use the Google Cloud SDK to retrieve data, upload predictions, and explain them.

Chapter 4

Case Studies

This chapter discusses the experimental process, datasets, and data preparation with all the relevant details on several use cases. The results on different datasets are compared against each other visually as well as through various metrics.

4.1 Evaluating parameters

The model is analysed with performance matrices as portrayed in Table [4.1, 4.2, 4.3, 4.4]. Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. Both precision and recall are therefore based on relevance. The definitions of precision and recall are:

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

The balanced F-score, often known as the classic F-measure, is a metric that is the harmonic mean of accuracy and recall:

$$F = \frac{precision \cdot recall}{precision + recall}$$

The quantity of true positives, false positives and false negatives is determined by confusion matrices. The quantity of inaccurately erroneous forecasts is measured by the false-negative rate. The model's performance is shown using a confusion matrix and ROC-AUC curve, as shown in Figures [4.1, 4.2, 4.3, 4.4].

A ROC curve, also known as a receiver operating characteristic curve, compares the true positive rate and the false positive rate at various classification levels. More items are classified as positive when the classification threshold is lowered, which raises the number of both False Positives and True Positives. The 'Area Under the ROC Curve', often known as AUC, is an overall indicator of performance overall potential categorization criteria. The likelihood that the model values a random positive example higher than a random negative example is one approach to analyse AUC.

4.2 Case I: Breast Cancer Wisconsin Diagnosis

Table 4.1: Classification report for 75:25 train-validation ratio of Breast Cancer Dataset

	precision	recall	f1-score	support
0.0	0.99	0.98	0.98	92.0
1.0	0.96	0.98	0.97	51.0
accuracy			0.98	143.0
macro avg	0.97	0.98	0.98	143.0
weighted avg	0.98	0.98	0.98	143.0

The greatest cause of cancer-related death among women worldwide and the malignancy with the highest rate of diagnosis is breast cancer [33]. According to Mubarak's epidemiological research, breast cancer, a particularly deadly kind of cancer, can cause death and mortality in women if it is not recognised in its early stages [34]. It can be found using a variety of techniques, including X-ray mammography, 3-D ultrasound, computed tomography, positron emission tomography, magnetic resonance imaging (MRI), and breast temperature monitoring, although a pathology diagnosis is the most reliable [35]. Sex, age, oestrogen, family history, gene abnormalities, an unhealthy lifestyle, and other variables linked to the development of the illness are only a few of the many risk factors for breast cancer [36].

Our dataset contains tabular data with 32 features and over 569 data points. A fine needle aspirate (FNA) of a breast lump is used to generate the features from a digital image in 3-dimension. The model is trained-tested with 75:25 ratio of this dataset. With

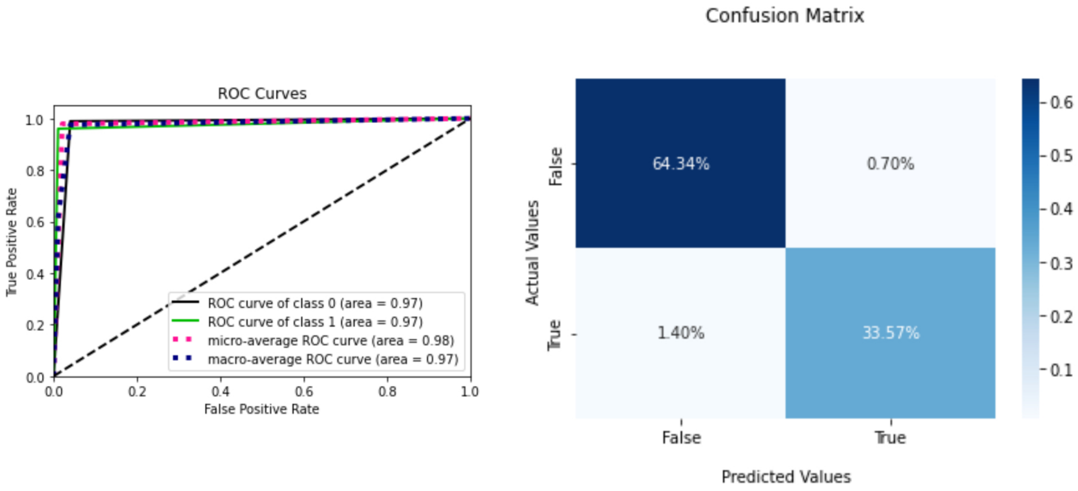

Figure 4.1: ROC curve and Confusion matrix for Breast Cancer Dataset

such specifications, AutoHealthX produced an accuracy of 98% which is much better than existing cloud-based models. Further, the Confusion matrix and the ROC curve is also shown in figure 4.1.

4.3 Case II: Heart Disease Cleveland Dataset

Table 4.2: Classification report for 75:25 train-validation ratio for Heart Disease Cleveland Dataset

	precision	recall	f1-score	support
0.0	0.85	0.75	0.78	40.0
1.0	0.78	0.86	0.80	39.0
accuracy			0.79	75.0
macro avg	0.79	0.79	0.79	75.0
weighted avg	0.80	0.79	0.79	75.0

Several different cardiac disorders are referred to as "heart disease." Coronary artery disease (CAD), which disrupts the blood flow to the heart, is the most typical kind of heart disease in the United States. A heart attack may result from reduced blood flow. Heart illness can sometimes go unnoticed until a person exhibits the early symptoms or signs of a cardiac arrest, heart failure, or an arrhythmia [37].

The dataset implemented on the AutoHealthX model constitutes over 300 patients' data with 75 attributes. With a ratio of 75:25 from this dataset, the model is trained and

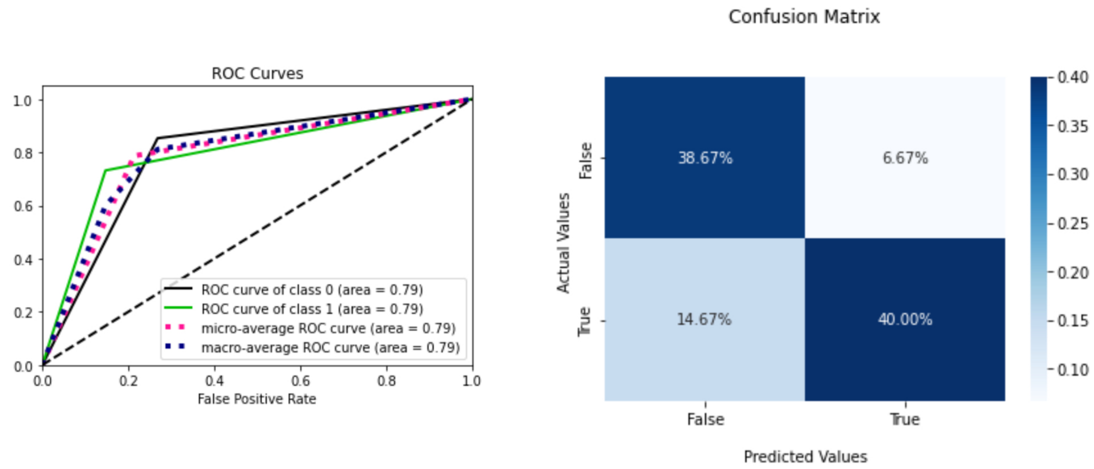

FIGURE 4.2: ROC curve and Confusion matrix for Heart Disease Cleveland Dataset

validated. With these conditions, AutoHealthX obtained an accuracy of 79% with a precision and recall of 0.85 and 0.75 respectively(as in table 4.2). In addition, figure 4.2 also displays the ROC curve and the Confusion matrix.

4.4 Case III: Diabetes Dataset

TABLE 4.3: Classification report for 75:25 train-validation ratio for Diabetes Dataset

	precision	recall	f1-score	support
0.0	0.92	0.81	0.86	148.0
1.0	0.85	0.80	0.83	51.0
accuracy			0.82	192.0
macro avg	0.83	0.86	0.84	192.0
weighted avg	0.83	0.88	0.80	192.0

Diabetes is a chronic condition brought on by either insufficient insulin production by the pancreas or inefficient insulin use by the body. The hormone called insulin controls blood sugar levels. Uncontrolled diabetes frequently results in hyperglycemia, or elevated blood sugar, which over time causes substantial harm to many different bodily systems, including the neurons and blood vessels. A total of 1.5 million fatalities were directly related to diabetes in 2019, and 48% of these deaths occurred in those under the age of 70. Diabetes contributed to an additional 460 000 renal disease deaths, and high blood glucose is responsible for 20% of cardiovascular fatalities around the world [38].

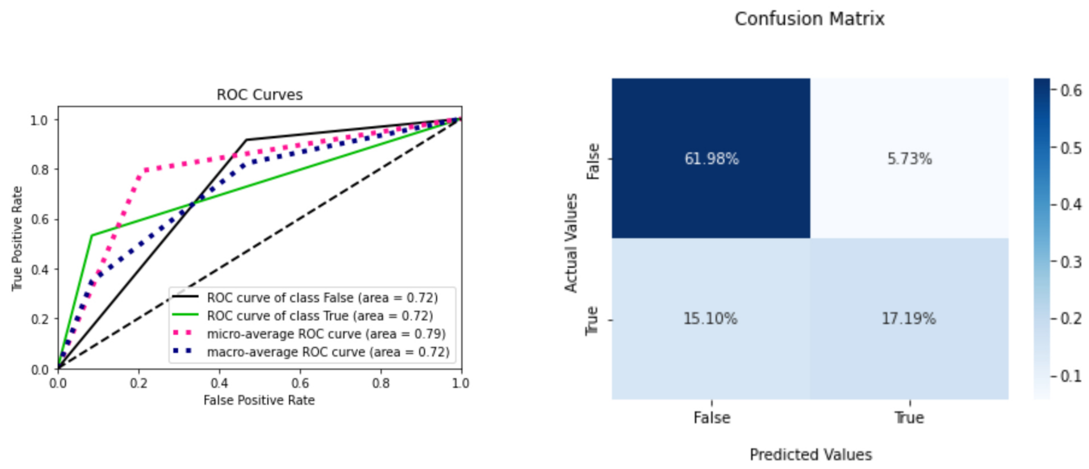

FIGURE 4.3: ROC curve and Confusion matrix for Diabetes Dataset

To test the AutoHealthX, the Diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Disease is taken into consideration. The dataset constituted 10 characteristics and 768 instances with all patients being Pima Indian females who are at least 21 years old. The model is trained-tested on the 75:25 ratio of the dataset and achieved an overall accuracy of 82%. The Classification report is provided in table 4.3. Further, the ROC-AUC curve and confusion matrix is given in figure 4.3.

4.5 Case IV: COVID-19 Dataset

TABLE 4.4: Classification report for 75:25 train-validation ratio

	precision	recall	f1-score	support
0.0	0.90	0.92	0.91	148.0
1.0	0.85	0.80	0.83	51.0
accuracy			0.86	200.0
macro avg	0.83	0.82	0.82	200.0
weighted avg	0.86	0.86	0.86	200.0

In general, human life and health have been profoundly damaged by the SARS Cov2-led COVID-19 pandemic. The majority of COVID-19 patients have mild to moderate symptoms. However, this devastating outbreak caused suffering and death in people. COVID-19 has the propensity to target and harm lung tissue. This devastating outbreak

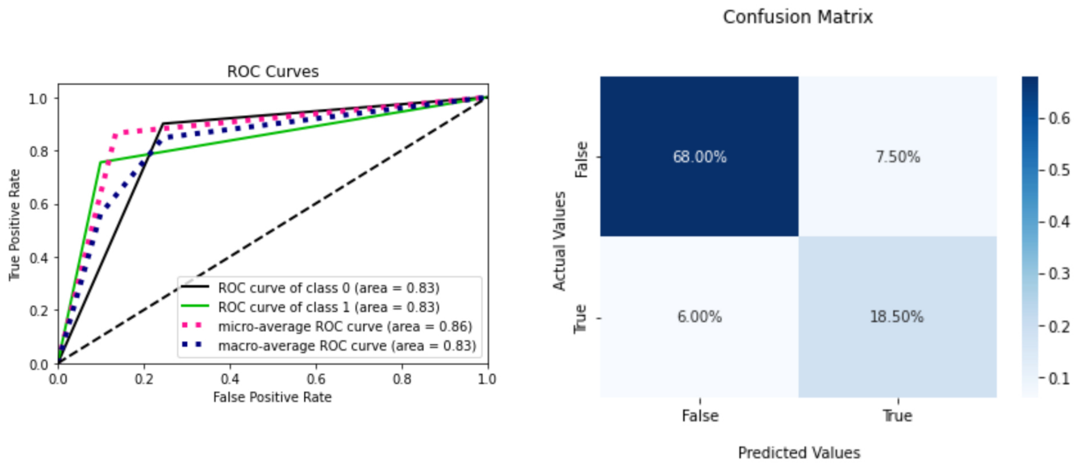

Figure 4.4: ROC curve and Confusion matrix for COVID-19 Dataset

caused death in 6,657,706 people around the world. Because of its fast-spreading ability, the World Health Organization (WHO) designated COVID-19 a Public Health Emergency.

For Covid-19 dataset, tabular data is used with about 800 patient data. It contains 26 attributes such as age, heart conditions, smoking, pregnancy, etc. With a ratio of 75:25 from this dataset, the model is trained and validated. With these conditions, AutoHealthX obtained an accuracy of 86% as shown in table 4.4. In addition, figure 4.4 also displays the ROC curve and the Confusion matrix.

Chapter 5

Web Application

The aim of developing AutoHealthX is to make AI usable for healthcare professionals and normal users, without any technical knowledge. To make it possible, an interractive web application is developed using StreamLit Framework. The Application is deployed on StreamLit cloud and Google App Engine.

5.1 Working Principle:

The dataset is uploaded by the user into a cloud storage bucket using the web interface [5.1].

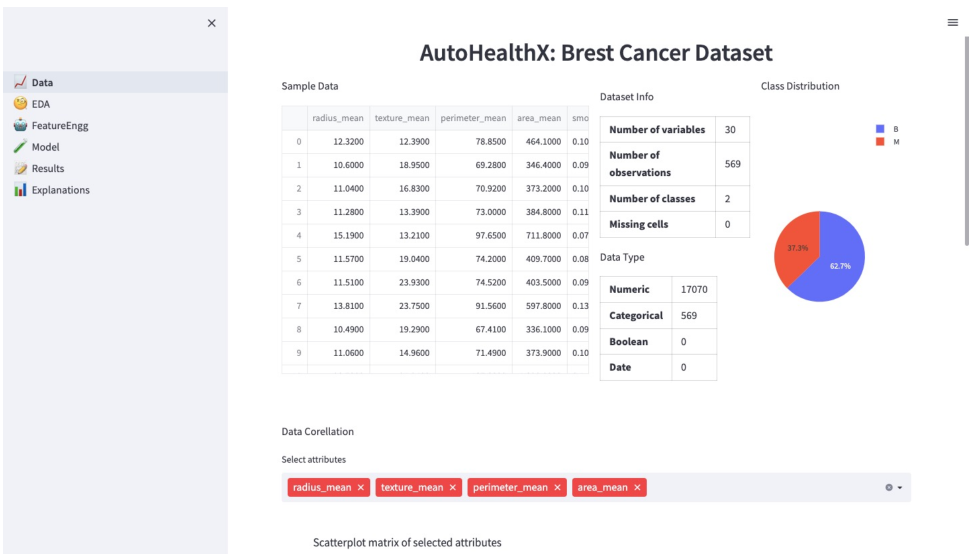

FIGURE 5.1: The Dataset page with Breast Cancer Dataset

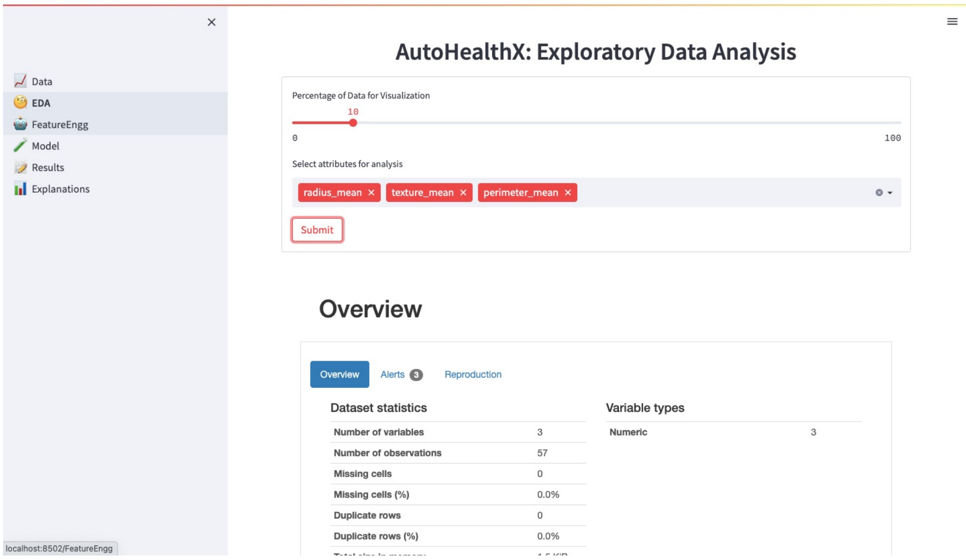

FIGURE 5.2: The EDA page with Breast Cancer Dataset

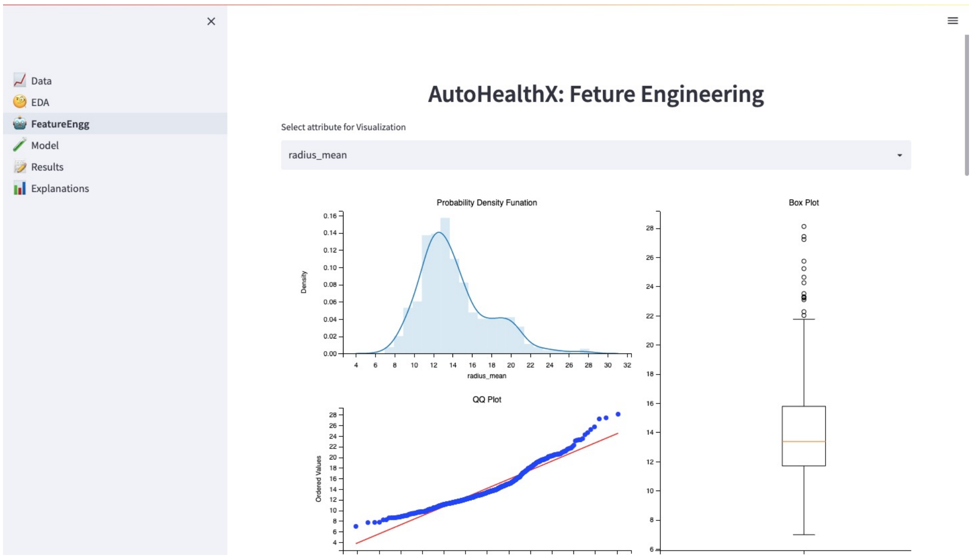

FIGURE 5.3: The Feature Engineering page with Breast Cancer Dataset

Then, User can perform Exploratory Data Analysis over the dataset using Pandas Profiling [5.2]. Moreover, the feature engineering cloud function gets triggered and does the feature engineering part automatically [5.3]. Users have the power to fine-tune the FE part manually as well using the dashboard. After FE, AutoML cloud function comes to the picture which implements the appropriate ML algorithm using Auto-Keras [5.4]. The same function generates the training and testing dataset along with confusion matrics and and classification report for results. Lastly, LIME explainer cloud function generates 5 sample explaination which is then displayed into the user screen in an HTML format[5.5].

Figure 5.4: The Results page with Breast Cancer Dataset

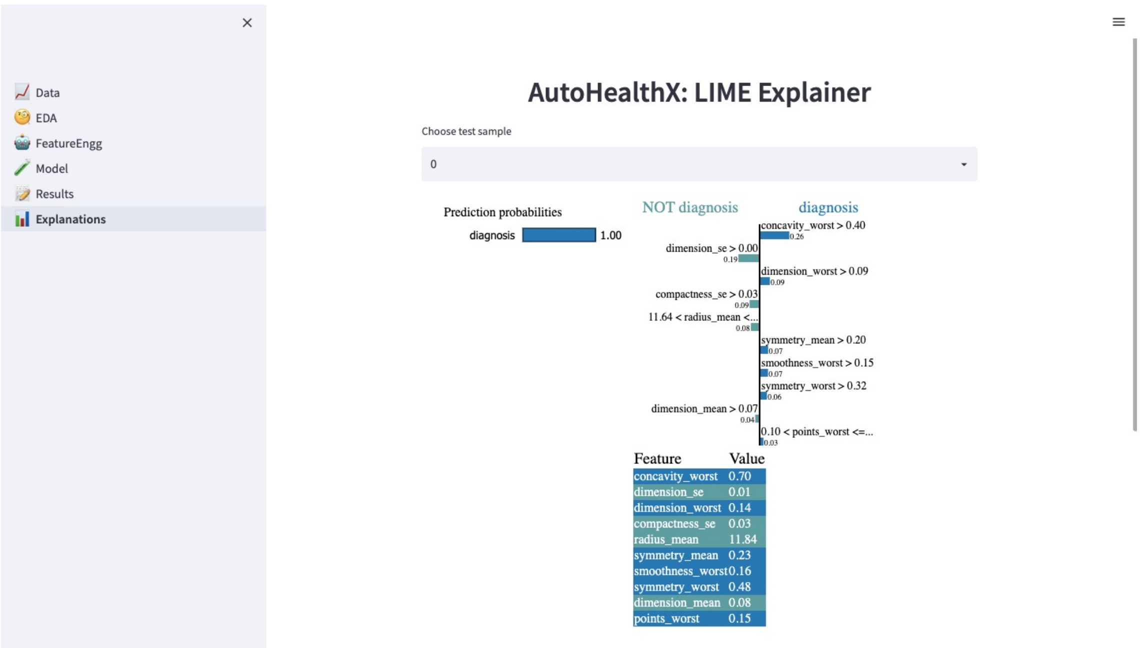

Figure 5.5: The LIME Explainer page with Breast Cancer Dataset

Chapter 6

Final Remarks

6.1 Discussion

By testing several instances of the function in different conditions, we have realized that the execution lasted almost 2 minutes on average. Although there are instances when the 'auto_ml' cloud function executes for up to 4 minutes, it's evident that the large dataset requires more memory and hardware capacity. On the other hand, we are also able to understand from the figure6.1 that many of the function instances have crashed due to several reasons. Two of the major reasons are the incompatibility of the dataset and the long execution time. As GCP has a limit on the execution time of up to 5 minutes for cloud functions, the implementation of a large ML model with a long training time might be problematic. There are multiple ways to fix this issue, which include using multiple functions parallelly or consecutively for a long training period.

Not only does execution time matter for cloud functions, but the memory capacity of the functions also plays a very important role in a cloud environment. According to figure 6.2, most of the function execution took 500MB to 1000MB of memory bandwidth. The

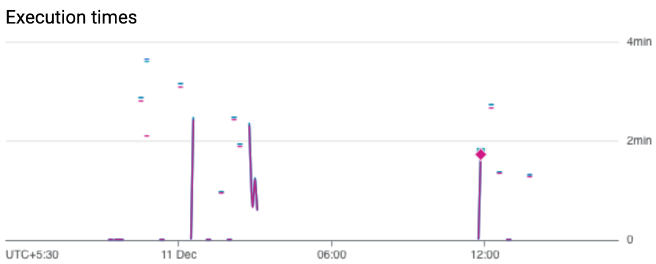

FIGURE 6.1: Execution Time Graph of 'auto_ml' cloud function

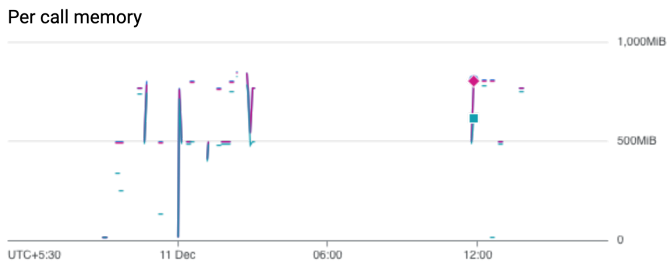

FIGURE 6.2: Per call memory utilization Graph of 'auto_ml' cloud function

right balance of memory and processing capability is crucial for smooth function execution. As these are tabular datasets, this execution requires less amount of memory bandwidth in comparison to the image and time series datasets. That means while dealing with image and time-series data we must increase the function memory capacity. The memory bandwidth of the 'auto_ml' function is set at 4GB at max so that any interruption can be prevented.

6.2 Conclusion

This study developed a novel framework for automated machine learning. The goal of creating AutoHealthX is to make AI accessible to non-technical individuals and healthcare professionals. It is made feasible by utilising the StreamLit Framework to create an interactive web application. Both Google App Engine and the StreamLit cloud are used to deploy the application. The model overall achieved a maximum accuracy of 98%.

6.3 Future Works

1. The model is primarily addressing the classification data problem as it is the most prominent use case in the healthcare domain. It can be extended into regression problems as well.

2. StrutureDataClassifier and StrutureDataRegressor deal with tabular data and it doesn't count on the other forms of data such as time-series data. Although time-series data are less frequent than image data in the context of Heath care, it is useful while measuring real-time patient activity. The project can be extended in such a context.

3. Different forms of data like Images and video needs more processing capability which can be implemented using fog architecture. Which can be extended using a cloud model training schedule.

4. Real-time disease detection using on-device model predication can be implemented in edge-fog and cloud models.

5. Federated learning can be implemented for the privacy protection of the patients by which the learning model can improve from individual patients' model feedback.

Appendix A

The integrated development environment used for the coding of the model, pre and post-transfer learning, is Kaggle with features as follows:

- CPU: Kaggle kernel
- GPU: 15.9 GB
- RAM: 13 GB
- Disk: 73.1 GB
- Site: https://www.kaggle.com

Google Colab has been used for training purposes, with features such as:

- CPU: Python3 Google Compute Engine backend
- GPU: 15.9 GB
- RAM: 12.68 GB
- Disk: 107.72 GB
- Site: http://colab.research.google.com

Bibliography

[1] E. Brynjolfsson, L. M. Hitt, and H. H. Kim, "Strength in numbers: How does data-driven decisionmaking affect firm performance?," *Available at SSRN 1819486*, 2011.

[2] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 5–22, Springer, 2019.

[3] Y. Bengio, Y. LeCun, *et al.*, "Scaling learning algorithms towards ai," *Large-scale kernel machines*, vol. 34, no. 5, pp. 1–41, 2007.

[4] E. Lindholm, J. Nickolls, S. Oberman, and J. Montrym, "Nvidia tesla: A unified graphics and computing architecture," *IEEE micro*, vol. 28, no. 2, pp. 39–55, 2008.

[5] M. Garouani, A. Ahmad, M. Bouneffa, M. Hamlich, G. Bourguin, and A. Lewandowski, "Towards big industrial data mining through explainable automated machine learning," *The International Journal of Advanced Manufacturing Technology*, vol. 120, no. 1, pp. 1169–1188, 2022.

[6] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in *CCF international conference on natural language processing and Chinese computing*, pp. 563–574, Springer, 2019.

[7] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," *Advances in neural information processing systems*, vol. 28, 2015.

[8] R. S. Olson and J. H. Moore, "Tpot: A tree-based pipeline optimization tool for automating machine learning," in *Workshop on automatic machine learning*, pp. 66–74, PMLR, 2016.

[9] L. Kotthoff, C. Thornton, H. H. Hoos, F. Hutter, and K. Leyton-Brown, "Auto-weka: Automatic model selection and hyperparameter optimization in weka," in *Automated machine learning*, pp. 81–95, Springer, Cham, 2019.

[10] T. Swearingen, W. Drevo, B. Cyphers, A. Cuesta-Infante, A. Ross, and K. Veeramachaneni, "Atm: A distributed, collaborative, scalable system for automated machine learning," in *2017 IEEE international conference on big data (big data)*, pp. 151–162, IEEE, 2017.

[11] U. Khurana, H. Samulowitz, and D. Turaga, "Feature engineering for predictive modeling using reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[12] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. S. Turaga, "Learning feature engineering for classification.," in *Ijcai*, pp. 2529–2535, 2017.

[13] M. Reif, F. Shafait, M. Goldstein, T. Breuel, and A. Dengel, "Automatic classifier selection for non-experts," *Pattern Analysis and Applications*, vol. 17, no. 1, pp. 83–96, 2014.

[14] R. Vainshtein, A. Greenstein-Messica, G. Katz, B. Shapira, and L. Rokach, "A hybrid approach for automatic model recommendation," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1623–1626, 2018.

[15] M. Feurer, J. Springenberg, and F. Hutter, "Initializing bayesian hyperparameter optimization via meta-learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[16] B. Bilalli, A. Abelló, T. Aluja-Banet, and R. Wrembel, "Automated data pre-processing via meta-learning," in *International Conference on Model and Data Engineering*, pp. 194–208, Springer, 2016.

[17] B. Bilalli, A. Abelló, T. Aluja-Banet, R. F. Munir, and R. Wrembel, "Presistant: data pre-processing assistant," in *International Conference on Advanced Information Systems Engineering*, pp. 57–65, Springer, 2018.

[18] I. Guyon, L. Sun-Hosoya, M. Boullé, H. J. Escalante, S. Escalera, Z. Liu, D. Jajetic, B. Ray, M. Saeed, M. Sebag, *et al.*, "Analysis of the automl challenge series," *Automated Machine Learning*, p. 177, 2019.

[19] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[20] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *International conference on learning and intelligent optimization*, pp. 507–523, Springer, 2011.

[21] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012.

[22] K. Tu, J. Ma, P. Cui, J. Pei, and W. Zhu, "Autone: Hyperparameter optimization for massive network embedding," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 216–225, 2019.

[23] B. Hall and D. D. Henningsen, "Social facilitation and human–computer interaction," *Computers in human behavior*, vol. 24, no. 6, pp. 2965–2971, 2008.

[24] M. Garouani, A. Ahmad, M. Bouneffa, A. Lewandowski, G. Bourguin, and M. Hamlich, "Towards the automation of industrial data science: A meta-learning based approach.," in *ICEIS (1)*, pp. 709–716, 2021.

[25] "Breast cancer wisconsin (diagnostic) data set," 1995.

[26] K. P. Bennett and O. L. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization methods and software*, vol. 1, no. 1, pp. 23–34, 1992.

[27] U. M. L. Repository, "Heart disease data set," 1998.

[28] N. I. of Diabetes, Digestive, and K. Diseases, "Diabetes dataset," 2022.

[29] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J.-L. Suárez, J. Luengo, M. Valero-González, *et al.*, "Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images," *IEEE journal of biomedical and health informatics*, vol. 24, no. 12, pp. 3595–3605, 2020.

[30] D. Chowdhury, S. Poddar, S. Banarjee, R. Pal, A. Gani, C. Ellis, R. C. Arya, S. S. Gill, and S. Uhlig, "Covidxai: Explainable ai assisted web application for covid-19 vaccine prioritisation," *Internet Technology Letters*, p. e381, 2022.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[32] T. Peltola, "Local interpretable model-agnostic explanations of bayesian predictive models via kullback-leibler projections," *arXiv preprint arXiv:1810.02678*, 2018.

[33] L. Wilkinson and T. Gathani, "Understanding breast cancer as a global health concern," *The British Journal of Radiology*, vol. 95, no. 1130, p. 20211033, 2022.

[34] S. Mubarik, Y. Yu, F. Wang, S. S. Malik, X. Liu, M. Fawad, F. Shi, and C. Yu, "Epidemiological and sociodemographic transitions of female breast cancer incidence, death, case fatality and dalys in 21 world regions and globally, from 1990 to 2017: an age-period-cohort analysis," *Journal of advanced research*, vol. 37, pp. 185–196, 2022.

[35] X. Zhou, C. Li, M. M. Rahaman, Y. Yao, S. Ai, C. Sun, Q. Wang, Y. Zhang, M. Li, X. Li, *et al.*, "A comprehensive review for breast histopathology image analysis using classical and deep neural networks," *IEEE Access*, vol. 8, pp. 90931–90956, 2020.

[36] W. Majeed, B. Aslam, I. Javed, T. Khaliq, F. Muhammad, A. Ali, and A. Raza, "Breast cancer: major risk factors and recent developments in treatment," *Asian Pacific Journal of Cancer Prevention*, vol. 15, no. 8, pp. 3353–3358, 2014.

[37] "About heart disease," 2022.

[38] "Diabetes," 2022.

[39] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[40] S. Di Maria, S. Vedantham, and P. Vaz, "X-ray dosimetry in breast cancer screening: 2d and 3d mammography," *European Journal of Radiology*, p. 110278, 2022.

[41] S. Iftikhar, M. Golec, D. Chowdhury, S. S. Gill, and S. Uhlig, "Fogdlearner: A deep learning-based cardiac health diagnosis framework using fog computing," in *Australasian Computer Science Week 2022*, pp. 136–144, 2022.