Open Source Robust Machine Learning Software for Medical Patient Data Analysis and Cloud Storage

Md. Sakib Abrar Hossain and Md. Ashrafuzzaman(✉)

Department of Biomedical Engineering,
Military Institute of Science and Technology, Dhaka 1216, Bangladesh
blksakib@gmail.com, ashezaman@gmail.com

Abstract. Big data and artificial intelligence-based researches in the health care arena have radically changed the sector with better preventive health care, early diagnosis of diseases, and advanced assistive technology along with numerous other areas. Health care facilities, academic research centers, and industries are collaborating in developed countries on such researches. Besides, developing and underdeveloped countries stay behind in this field of research due to infirm health and e-health infrastructure, insufficient technical manpower, low physicians to patient ratio, and other limitations. Our research focuses on developing an open-source and easy to use Machine Learning Software System that should uplift Big Data and data science researches focusing on health care in the developing and underdeveloped countries amid such obstacles. This pilot study is a part of that big project that helps to make sense about the working methodology and the expected outcomes by the end of the project. Apart from medical data analysis, it could serve as an efficient platform for storing patient data and we hope academicians, professionals, and physicians around the globe will be aided by such robust data analysis software, as it facilitates automated preprocessing of data, building and comparing different prediction models, cloud storage and data visualization techniques. This work visualizes most of the part of its concept to understand its facilities, although due to some restriction some techniques will be discussed only after completing this big project.

Keywords: Health care informatics · Machine learning · Data visualization

Classification: Health care informatics · Artificial intelligence in healthcare

1 Introduction

The entire health care sector is currently experiencing a major shift due to recent breakthroughs in Artificial Intelligence. Artificial Intelligence now has entered the domain of human expertise in health care by completely transforming the medical-image diagnostic systems [1]. With the progress in Big Data and Machine Learning research, it has been observed that the combination of both of the fields can develop systems that can attain the same efficiency of human physicians [2].

T. Jarm et al. (Eds.): EMBEC 2020, IFMBE Proceedings 80, pp. 932–943, 2021.
https://doi.org/10.1007/978-3-030-64610-3_104

Since the late 2000 s Machine learning Researches focusing on health care has enormously increased. Figure 1 demonstrates the number of researches performed in the field of Machine Learning focusing on Health Informatics from 2008 until the year 2016 [3].

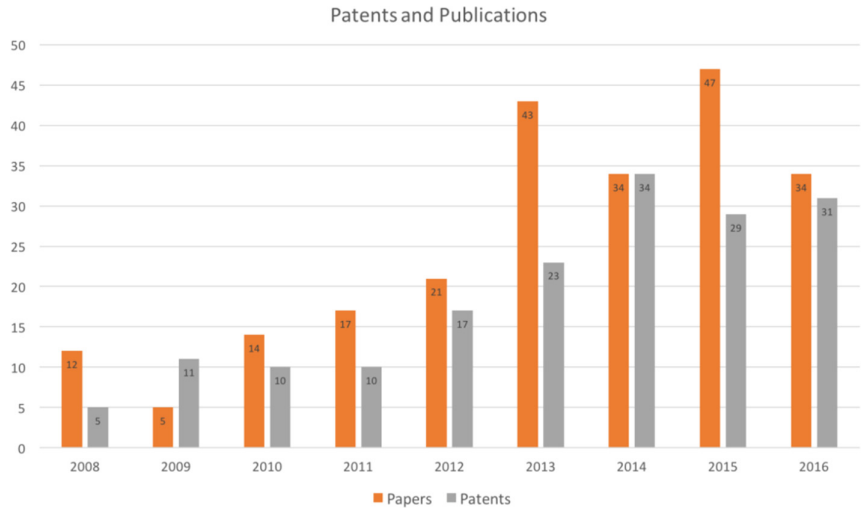

Fig. 1. Patents and research publications on machine learning in health informatics from the year 2008 to 2016.

Artificial Intelligence is now able to solve some critical issues in health care which even includes problems such as human resource crisis [4]. Currently, the global prime cause of death is cardiovascular disease, which represents 31% of total deaths around the globe [5]. Artificial Intelligence can provide efficient solutions to such problems by improving preventive healthcare with machine learning-based cardiovascular risk prediction models [6]. Along with many other areas, Artificial Intelligence is also playing a vital role in Assistive Technologies in health care [7, 8].

1.1 Research Regarding Developing & Underdeveloped Countries

Artificial Intelligence researches focusing on health care regarding underdeveloped and developing countries are going on for a while. Deep learning-based hospital management systems can play a vital role in the underdeveloped and developing countries, where there is a crisis for skilled health professionals [9]. Artificial Intelligence is also used to create cheap biosensors which can be effective in the context of developing and underdeveloped economies [10]. Telemedicine systems can play an important role in providing health care for the underprivileged. In developing countries like Bangladesh, such telemedicine systems have been functional for a long time, providing health care solutions in the underprivileged communities [11]. Artificial Intelligence integrated within the telemedicine system can grant a different dimension for providing healthcare in such communities [12]. Artificial Intelligence technologies can provide telediagnosis and teletreatment for underprivileged communities [13].

In the developed countries, the collaboration for artificial intelligence research focusing on healthcare between hospitals, health care facilities, universities, research institutes, and industries have made it possible for such improvement in this sector. But on the other hand, due to lower patient to doctor ratio [14] along with other factors, hospitals and health facilities in the underdeveloped and developing areas can not avail the scope of participating in such researches. Along with this, due to infirm e-health infrastructure, patient data is not always available in developing countries and underdeveloped countries. All these facts all together have made it difficult for collaborative research in the developing and underdeveloped countries.

1.2 The Barrier for Physicians in Data Science

Data Science requires professionals to have prior programming expertise. The absence of programming expertise prevents professionals to use different data visualization libraries, machine learning, and deep learning classifiers along with many other tools. Such an absence of expertise acts as a barrier to indirect involvement in data science researches for physicians.

1.3 Research Goal

The main goal of this research work is to develop open-source and user-friendly software, which will encourage the active participation of physicians in data science research. Additionally, the aim of the research also involves encouraging hospitals and other healthcare facilities in underdeveloped and developing countries for participating in data science research, by providing an open-source system for data analysis and storing.

2 Method

For developing the software python programming language has been chosen due to its robustness. The software is designed to solve and analyze classification based datasets, which belongs to the domain of supervised learning. For the training of models, the software uses multiple classifier algorithm from the scikit learn [15] python library. Python's pandas [16] library is used to generate data frames from input csv/excel formatted files. To keep the software lightweight, conventional databases like mysql, mongodb or others are not being used. For managing data pandas data frame along with python's unique dictionary data structure is used. For preprocessing of the data pandas library along with python's numpy [17] library is used. Seaborn [18], an advanced data visualization library for python along with matplotlib [19] are used for visualizing the data. For storing data files in cloud Google's firebase [20] service is used. For the graphical user interface of the software, tkinter package has been used.

2.1 Data Upload-Locally Storing-Preprocessing-Training

A pandas data frame is generated from the input CSV/Excel file. A copy of the data frame is stored locally for visualization and queries. Preprocessing of the data involves converting the data frame to multiple numpy objects, before checking for null or missing values. The dimension of the numpy objects is changed as per the parameters

of the models. Four classifiers are trained with Random Forest Classifier algorithm, Linear Discriminant Analysis algorithm, Gaussian Naïve Bias algorithm, and Extra Tree Classifier algorithm. The accuracy and run-time value of each model are stored in a dictionary. Figure 2 demonstrates the whole workflow of the proposed work.

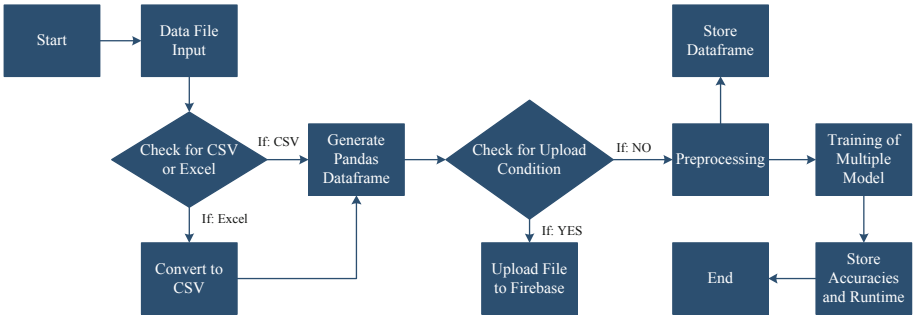

Fig. 2. Flow diagram for data upload, locally data storing, preprocessing and model training

Pre-processing: The pre-processing is a unique feature of this software. It performs null value and missing value checks on the data sets. Along with its feature scaling is performed. According to the feature value distribution the software chooses which type of scaling to be performed.

2.2 Generating Predictions

The input patient data is converted to a pandas data frame for reading. From the data frame, a numpy object is generated. From the trained models, the top-performing model is chosen (the model which has the largest accuracy value and the smallest run time value). The numpy object is then passed as the input to the selected model to generate the prediction. The whole process is demonstrated in the flow chart of Fig. 3.

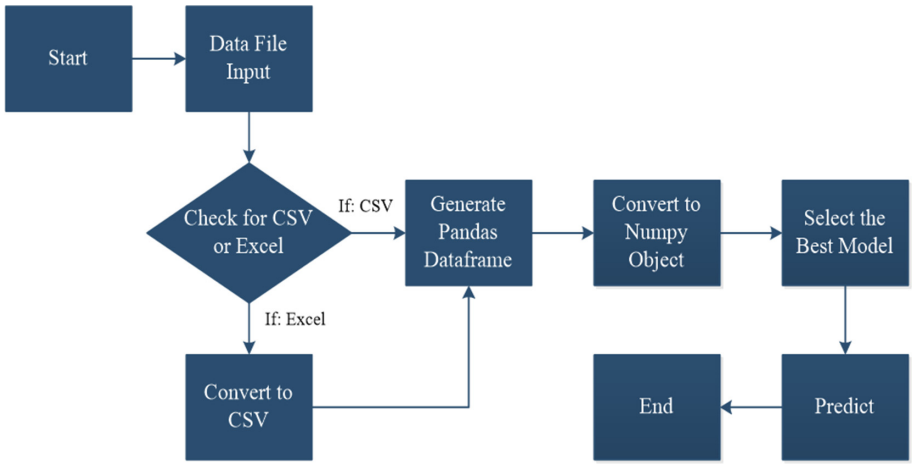

Fig. 3. Generating prediction

2.3 Queries

For keeping the software lightweight conventional databases have not been used. For storing data pandas data frame along with the combination of dictionary objects are used. Patient data are stored using pandas data frame. Searching within the data frame is performed for accessing arbitrary samples or attributes of samples. Searching within the data frame can be performed easily using the user interface. Information about the trained models is stored in a dictionary object. The dictionary object can be accessed from the user interface.

2.4 Data Visualization

The software has many advanced data visualization features. Data visualization features have enjoyed the main focus during the development of the software, as strong data visualization feature is essential for any efficient data analysis software. The software holds features of both 3D and 2D plotting. 3D plotting has been performed using a matplotlib 3D library. For the 2D plotting of data, the seaborn package has been used. The software enables the user for visualizing data through a Scatter plot, Box plot, Swamp Plot, Heat Map, and Parallel Coordinates in 2D.

Heat Map: The heat map shows the relation among the features in the dataset.
Parallel Coordinates: The parallel Coordinates plot shows how much separable the dataset is.
Swamp Plot: The Swamp plot generates each data point with its features in a unique 2D plot.
3d Plot: The 3D plot plots two given features in a three-dimensional space and shows there relation with the label values.

3 Result

The GitHub repository of this software [21, 22] has been kept open source. The repository contains a readme file that conveys instructions to run the software. The software has two versions, one for running on the Windows Operating system and the other for Linux Distros. For demonstrating the results a Breast Cancer Data set from the UCI machine learning repository [23] has been used. The key user interface of the designed software is illustrated in Fig. 4.

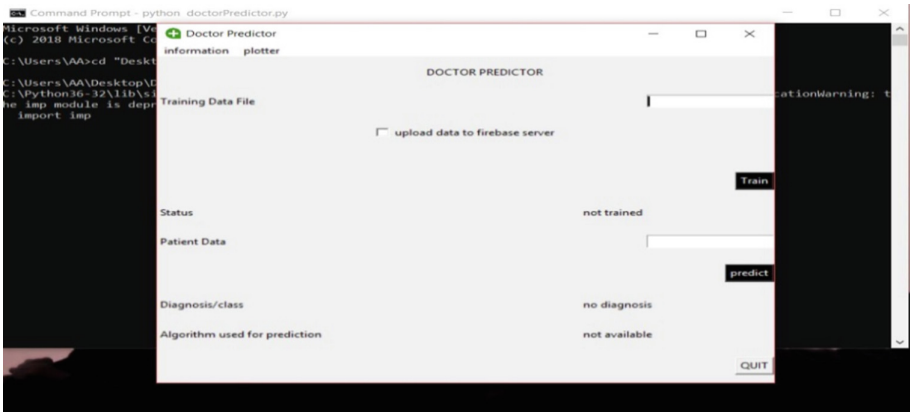

Fig. 4. The designed software and its key interface

Upon marking the checkbox the software uploads the data set in firebase for cloud storage (Fig. 5).

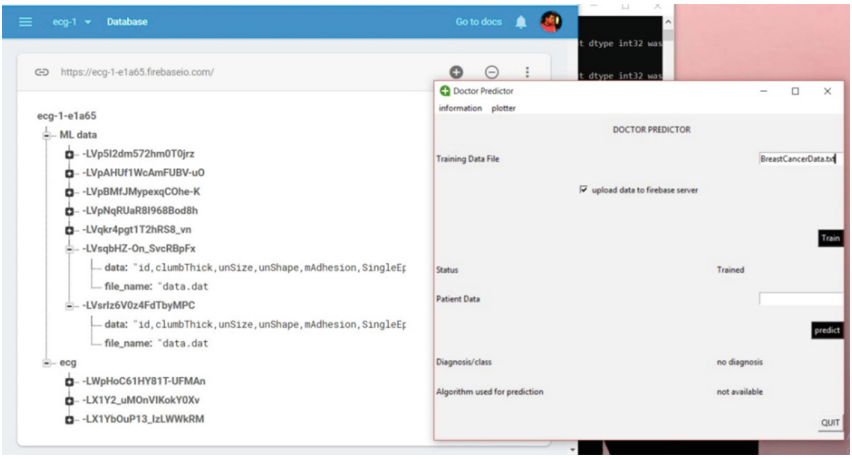

Fig. 5. Storing data in the cloud

Training of models, prediction, and accessing model information is shown in Fig. 6.

From the information menu in the upper left corner: model information, patient information, and feature/attribute information can be accessed. The patient information and feature information perform searches in a pandas data frame. The model information shows the values of a python dictionary (Fig. 7).

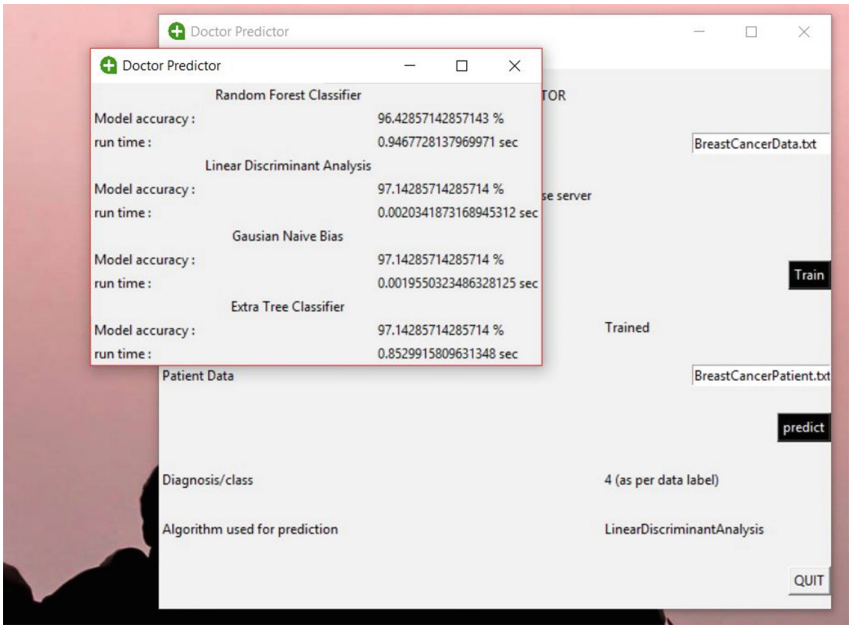

Fig. 6. Accessing model information, model training, and prediction.

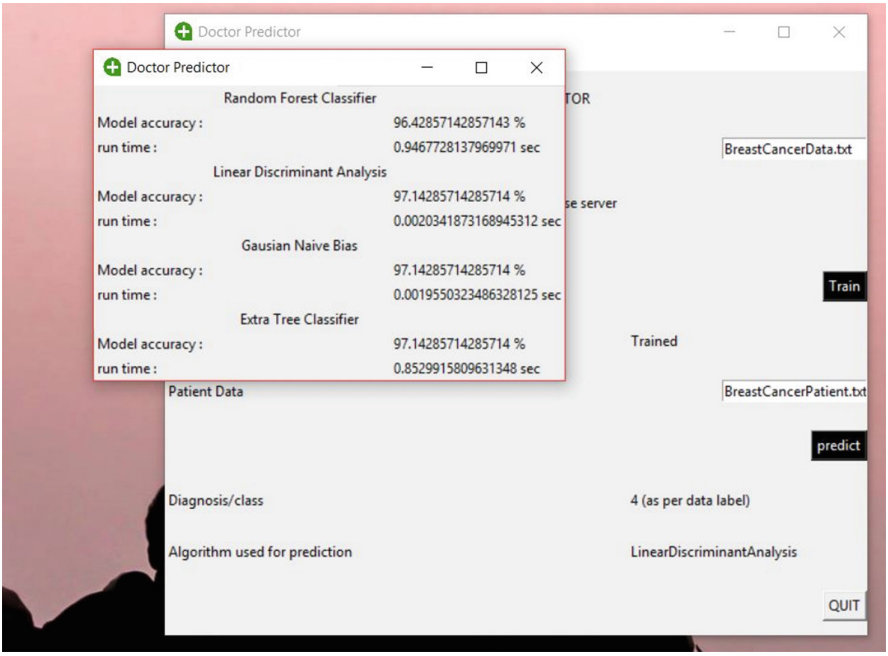

Fig. 7. Accessing the information menu

Users can select 2D or 3D plots from the plotter menu (located just beside the information menu). It shows submenus for 2D and 3D plotting options.

The 2D plotter has options for Scatter Plot, Box Plot, Heat Map, Parallel Coordinates, and swamp plot (Fig. 8).

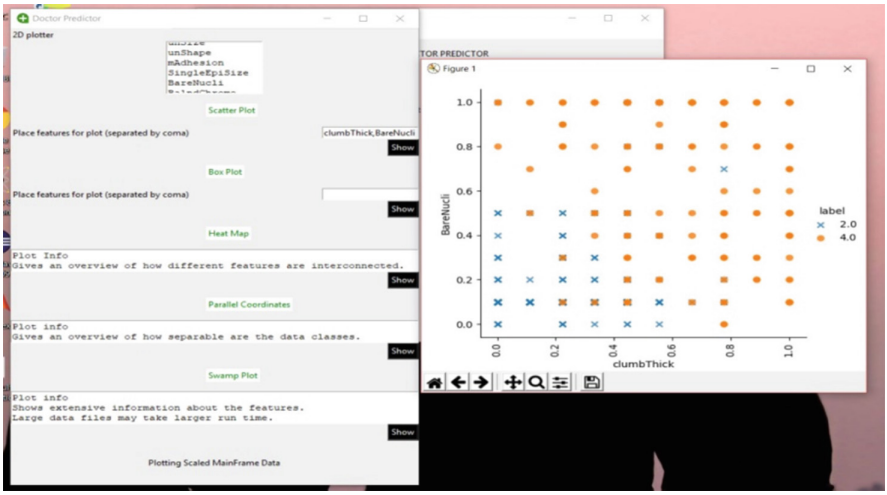

Fig. 8. 2D scatter plot

Both the Scatter Plot and Box plot takes two variables as inputs. Box plot represents feature values through their quartiles (Fig. 9).

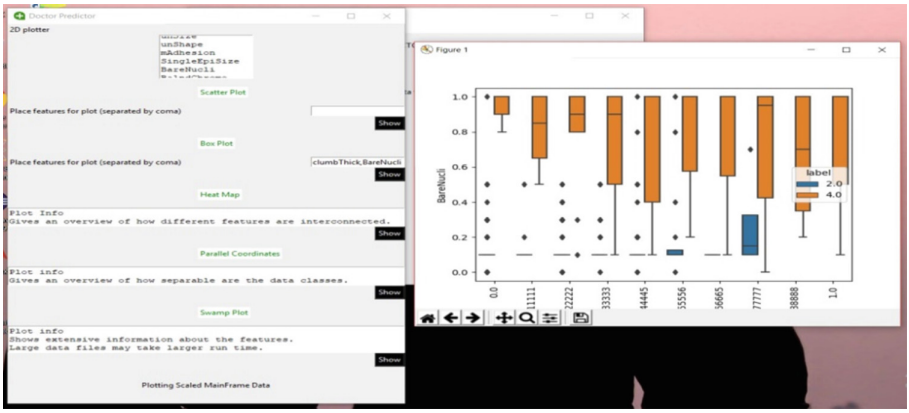

Fig. 9. 2D box plot

The Heat Map plot shows a correlation among the features. As discussed earlier the parallel Coordinate plot enables us to see how separable the data set is (Fig. 10).

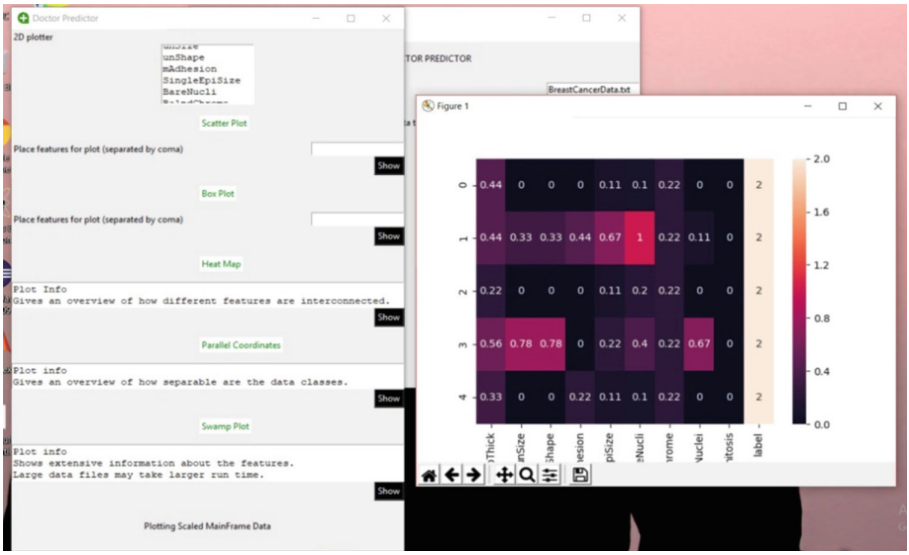

Fig. 10. 2D box plot

The swamp plot enables us more advanced visualization of the dataset, as it plots each point of the dataset. Due to the enormous amount of calculations that the computer has to perform during the swamp plot, performing swamp plots in the large dataset may take a longer time (Figs. 11, 12 and 13).

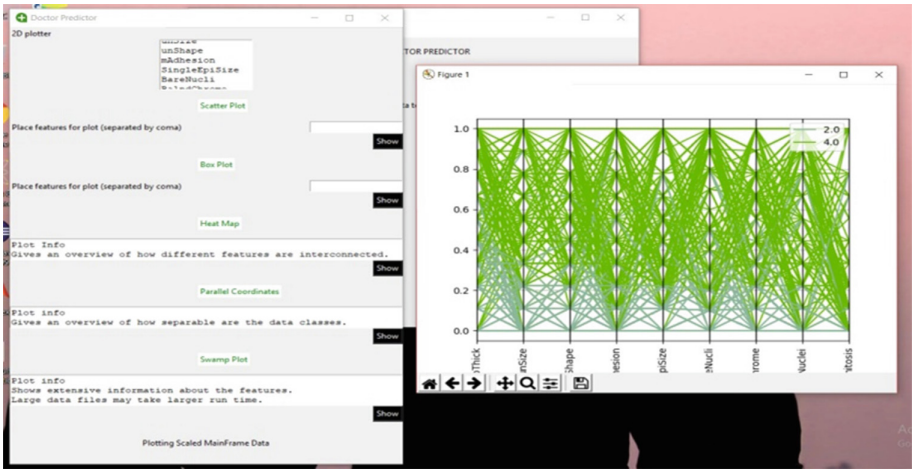

Fig. 11. 2D parallel coordinates plot

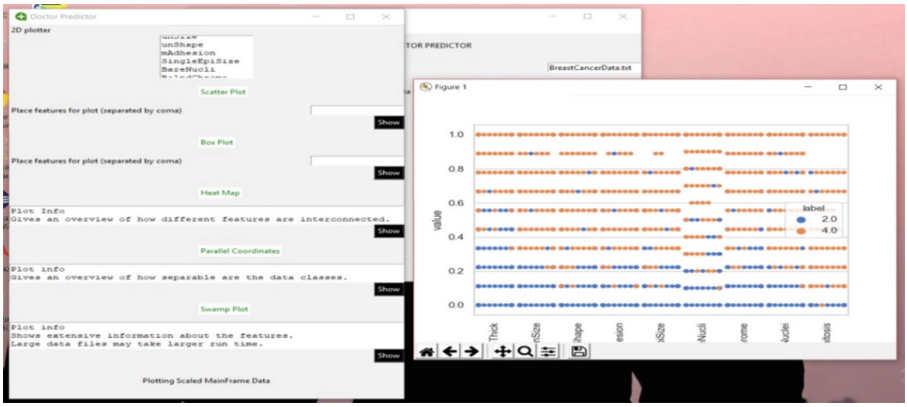

Fig. 12. Swamp plot

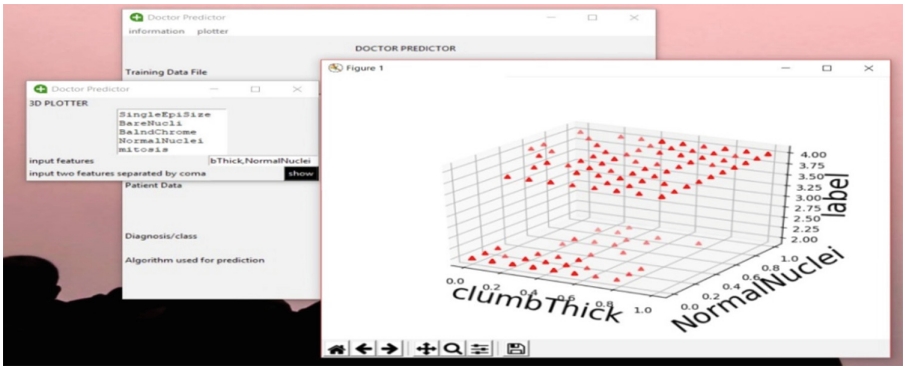

Fig. 13. 3D plot with two features

The 3D plot takes two features as parameters and plots the feature values in 3D.

4 Conclusion

The robustness of the developed Machine Learning software comes with advanced data visualization features. Along with this and basic features of model training, model comparison and predictions will give an edge to physicians without any programming expertise in pursuing data science researches. The feature of storing data in cloud storage will enable health professionals working in health facilities of the underdeveloped and developing countries by contributing huge amounts of medical data, which will be extremely helpful for carrying out data science researches in health care.

Economic growth is very closely related to improvement in health care, where data science researches are playing a crucial role [24]. We strongly believe such initiatives

of developing open-source software for data science will uplift health care researches, especially in underdeveloped countries, thus contributing to economic growth.

Lastly, Department of Biomedical Engineering at various universities and newly established Biomedical Engineering Institute in Bangladesh have been conducting researches focusing on the context of developing and underdeveloped health economies to ensure quality and improved healthcare services. Researches such as the development of a low-cost central monitoring platform [25], the establishment of a cardiovascular database are among such many initiatives. Involvement from WHO/IFBME/AAMI/IMDRF/IUPESM/AHWP/ACCE/ECRI/PAHO/JACE/CEASA will surely uplift both the enthusiasm and quality of researches in healthcare technology management filed of lower middle income countries globally.

Acknowledgment. The authors are greatly acknowledged by the support of the Department of Biomedical Engineering, Military Institute of Science and Technology.

Conflicts of Interest. The authors declare no conflict of interest regarding this research work.

References

1. Yu, K.-H., Beam, A., Kohane, I.: Artificial intelligence in healthcare. Nat. Biomed. Eng. **2**, 719–731 (2018). https://doi.org/10.1038/s41551-018-0305-z
2. Beam, A., Kohane, I.: Big data and machine learning in health care. JAMA **319**, 1317–1318 (2018). https://doi.org/10.1001/jama.2017.18391
3. Srivastava, S., Soman, S., Rai, A., Srivastava, P.K.: Deep learning for health informatics: recent trends and future directions. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1665–1670 (2017)
4. Mesko, B., Hetényi, G., Gyorffy, Z.: Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv. Res. **18**, 545 (2018). https://doi.org/10.1186/s12913-018-3359-4
5. World Health Organization. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
6. Weng, S., Reps, J., Kai, J., Garibaldi, J., Qureshi, N.: Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE **12**, e0174944 (2017). https://doi.org/10.1371/journal.pone.0174944
7. Bharucha, A., Anand, V., Forlizzi, J., Dew, M., Reynolds, C., Stevens, S., Wactlar, H.: Intelligent assistive technology applications to dementia care: current capabilities, limitations, and future challenges. Am. J. Geriatr. Psychiatry: Official J. Am. Assoc. Geriatr. Psychiatry **17**, 88–104 (2008). https://doi.org/10.1097/JGP.0b013e318187dde5
8. Hoey, J., Boutilier, C., Poupart, P., Olivier, P., Monk, A., Mihailidis, A.: People, sensors, decisions: customizable and adaptive technologies for assistance in healthcare. ACM Trans. Interact. Intell. Syst. **2**, 20:1–20:36 (2013). https://doi.org/10.1145/2395123.2395125
9. Safari, G., Majidi, B., Khanzadi, P., Manzuri, M.: Cross-platform e-management for smart care facilities using deep interpretation of patient surveillance videos, pp. 1–6 (2018). https://doi.org/10.1109/icbme.2018.8703545
10. Vashistha, R., Dangi, A., Kumar, A., Chhabra, D., Shukla, P.: Futuristic biosensors for cardiac health care: an artificial intelligence approach. 3 Biotech **8**, 358 (2018). https://doi.org/10.1007/s13205-018-1368-y

11. Rabbani, K., Amin, A.A., Abir, R., Bodiuzzaman, A., Khan, A., Tarafdar, Z.: A rural health monitor with telemedicine. In: Biomedical Engineering (2011)
12. De Araújo Novaes, M.: Telecare within different specialties. In: Fundamentals of Telemedicine and Telehealth, pp. 185–254 (2020)
13. Qazi, S., Tanveer, K., ElBahnasy, K., Raza, K.: From telediagnosis to teletreatment. In: Telemedicine Technologies, pp. 153–169 (2019)
14. World Bank Data for "Physicians per 1,000 people". https://data.worldbank.org/indicator/SH.MED.PHYS.ZS?type=points&view=map
15. Scikit learn documentation. https://scikit-learn.org/stable/
16. Pandas documentation. https://pandas.pydata.org/pandas-docs/stable/
17. Numpy documentation. https://numpy.org/doc/
18. Seaborn documentation. https://seaborn.pydata.org/
19. Matplotlib documentation. https://matplotlib.org/contents.html
20. Firebase documentation. https://firebase.google.com/docs/
21. Software repository for windows. https://github.com/Remian/Doctor-Predictor
22. Software repository for Linux Distros. https://github.com/Remian/doctorPredictorLinux
23. UCI Machine Learning Repository for Breast Cancer Patients. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
24. Abedjan, Z., et al.: Data science in healthcare: benefits, challenges and opportunities. In: Consoli, S., Reforgiato, R.D., Petković, M. (eds.) Data Science for Healthcare. Springer, Cham (2019)
25. Ashrafuzzaman, Md., Hossain, Md., Joaa, A., Aziz, Md.: Development of low cost central monitoring platform by modeling and simulation for patients care in low middle income countries (2019). https://doi.org/10.1007/978-981-10-9035-6_136