Predictions And Analytics In Healthcare: Advancements In Machine Learning

Vangala Surya Teja Reddy[1], Gatla Akanksh[2], Satyam[3], Kalpana Kumari[4], Bhanu Talwar[5]

[1234]Undergraduate student, Lovely Professional University, Jalandhar, Punjab
[5]Professor, Lovely Professional University, Jalandhar, Punjab

---

Abstract - *As everyone in this world knows how technology is advancing every day and the drastic changes are occurring in every sector these days and so does in the healthcare sector. Many revolutions are coming our way from this sector because the integrated technology in this sector is helping the scientists, researchers, doctors, etc. to reach their target goals with the help of computers. They can get the exact match or near to that with the help of technology and in the same way the user can also be informed and be filled with the knowledge with the help of these technological advancements. Like, in today's world, we are seeing a lot of websites or apps that are helping users know their diseases by getting some input from them and they are also getting proper medication based on that result. In this review, there is information about predictive analytics that is framing the healthcare sector and making it more comfortable for the end-user to diagnose their diseases and take charge of them in minimal time. This paper is going to talk about various advancements that can lead to, what we say as, informed user experience, which will keep them informed and also, they will be able to cure themselves at an appropriate time.*

Key Words*: *Healthcare sector, technological advancements, diseases, predictive analytics, diagnose, user experience.

INTRODUCTION

It is very essential for the diagnosis of chronic diseases in the medical field as these diseases persist for a long time. Some of the most common chronic diseases include diabetes, strokes, heart disease, arthritis, cancer, hepatitis C. If we can detect the disease in an early phase, it will help in taking preventive actions and effective treatment at an initial stage which is always found helpful for patients[1] in curing their diseases.

There has been lots and lots of data that can help the healthcare sector to improve on many weak points and also to do different advancements through the use of modern technologies such as machine learning, artificial intelligence, data mining techniques, and many more. These latest technologies combined can open many new doors for the people in this world and can help scientists and doctors to reach a particular milestone in a very short period. There is a lot of potential that needs to be unlocked because the things which were to be done in years are now possible within a fraction of seconds if all the things are placed in order. So basically, it can improve the efficiency of the human brain and also be a part of a tremendous change of era with the use of technology. In this research paper, a problem that needs to be solved is presented so that further development in this area can be made, and also with the help of existing technologies, something can be improved by using a tremendous amount of data that is available to us. There is always a question in people's minds about how a particular disease occurs and how to be able to cure that in the first place. So, by appropriately using some technology, we got to know that now people can assure their well-being by sitting at their homes, and also if they got to know something bad about them, then they can cure it as soon as they know about it. We are going to predict some of the diseases for them to let them know whether they are having any chances of catching that disease. Moreover, users' data will be analyzed about various implications or factors related to getting some diseases at a particular time, age, etc.

LITERATURE SURVEY

As we think it is easy to predict or analyze data, but it is not, it is certainly not so easy to analyze big-data data[4] within it and there are too many challenges to predict something to the end user in the right way.

The difficulty of analyzing big data comes from its three dimensions, namely, variability, speed, and volume. 'Variety' suggests that large amounts of data are made up of many types of data, both formal and informal, e.g., doctor notes. Healthcare data is presented in a variety of formats and presentations from a variety of sources including (1) Clinical data on Electronic Health Records, medical imaging, machine-sensory data, and genomics, (2) clinical and R&D data e.g., from clinical trials. and journal articles, (3) job claims and cost data from health care providers and insurance companies, and (4) patient behavior and emotional data from wearable devices and communication posts, including Twitter feeds, blogs, Facebook status updates, healthcare communities, and web pages. 'Speed' refers to large data that is transmitted and obtained in real-time, e.g., critical signals, and usually arrives at an explosive rather than continuous level. 'Volume' means large data, in the name itself, is extremely large. For example, 3D CAT scanning usually takes 1 GB while one human genome takes about 3 GB of data[2]. The role of professionals in the field of health care should maintain a balance between business and technology, which paves the way for Big Data in healthcare information systems. Using Big Data in a health care environment has led to changes in the storage and handling of complex, randomized data sets[3].

Several researchers have been committed in recent years to evaluating the classification precision of various classification algorithms used in the Cleveland heart disease database [9], which is freely available through the University of California's online data mining repository. Many researchers have used this database since its inception to investigate various classification challenges using various classification algorithms. Detrano [10] deployed a logistic regression approach and achieved a classification accuracy of 77%.

The author of [11] researched on the Cleveland data set with the goal of analyzing global computational intelligence techniques and found that applying a new approach improved prediction performance. The performance of his proposed technique, on the other hand, is dependent on the qualities chosen by the algorithm.

P. Saranya's paper[7] has mentioned a different algorithm which is used for predicting accurate results after diagnosis. Machine decision diagnosis auxiliary algorithm is used to diagnose cancer with large datasets, this algorithm has 77% accuracy in predicting the disease. An online contextual learning algorithm is used because it minimizes the false positive rate in breast cancer screening diagnosis decisions. Also, Multiple Streams of 2D convolution networks are used for detecting false positives in the scan. When 3D Convolutional Neural Network is used it performs better as it has a sensitivity of 85% to 90% in detecting false positives. Gamification, Convolutional Neural Network algorithm, and Online contextual learning algorithm is used for predicting outputs as accurately as possible.

Dhiraj Dahiwade, Gajanan Patle, and Ektaa Meshram proposed a generic disease prediction system[8] based on the patient's symptoms. For disease prediction, researchers used the K-Nearest Neighbor (KNN) and Convolutional Neural Network (CNN) machine learning algorithms. The CNN algorithm has an accuracy of 84.5 percent in general disease prediction, which is higher than the KNN approach. In addition, KNN has a higher time and memory need than CNN. After predicting general disease, their system can calculate the risk of general disease, indicating whether the risk is lower or higher.

Min Chen, Yixue Hao, Kai Hwang, Lu Wang, Lin Wang, In this proposed paper[12], They streamline techniques for effective chronic disease outbreak prediction in disease-prone communities. They modified prediction models using real-world hospital data from central China between 2013 and 2015. They employed a latent component model to recreate the missing data to overcome the challenge of incomplete data. And experimented on a regional chronic disease of cerebral infarction. Using structured and unstructured data from hospitals, they suggested a new convolutional neural network (CNN)-based multimodal disease risk prediction algorithm. When compared to various common prediction algorithms, their new approach has a prediction accuracy of 94.8% and a convergence speed that is faster than the CNN-based unimodal disease risk prediction algorithm.

TECHNOLOGY USED

Various technologies like AWS, EC instance; tableau, Flask, Pandas, Numpy, and Scikit Learn have been used to implement this project and to analyze and predict the values of the users, the data is converted into meaningful information to let users experience the advancements in technology.

In this implementation to predict and analyze the data, different libraries of python(like pandas, NumPy, Scikit Learn) are being used for data processing, data visualization, and model building.

Numpy and Pandas have libraries that are used for any scientific computation, here we used this due to their intuitive syntax and high-performance computation capabilities.

Numpy: It is used to provide support for large multidimensional array objects like shape manipulation, logical and mathematical operation, sorting, selecting, Basic algebra & statistical operations, and much more. In Numpy we can perform element-wise operations on your dataset. It provides us with an enormous range of fast and efficient ways of manipulating data inside them. Numpy uses less memory to store data and work on arrays, which is very convenient to use. During Data processing we use Numpy's function like sort() for Sorting elements, to add elements we use concatenate() function. We can do indexing and slicing of the data according to our requirements.

Pandas: It is an open-source python package that is mostly used for data analysis and machine learning tasks. Pandas allow us to do time-consuming and repetitive work in a very simple way. This work includes Data cleaning, data fill, Data normalization, data visualization, statistical analysis, Data inspection, loading & saving data, and many more. Pandas are used as one of the most important data manipulation and analysis tools. Pandas allow importing from different file formats such as comma-separated-values, JSON, SQL database tables, and Microsoft Excel.

Scikit-learn: It is the most useful and robust library of python used for machine learning. Scikit-learn allows us to define ML algorithms and evaluate many other algorithms against one another. It also includes tools to help us preprocess our datasets. This package provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, etc. The real power of Scikit-Learn lies in its model evaluation and selection framework, where we can cross-validate the parameters of our models. With the help of this library, we can create a machine learning model that will predict the positive or negative output model. This library is focused on modeling data with the help of Ensemble methods, Feature extraction, feature selection & Supervised models.

Flask: Flask is a web framework of Python modules that let us develop web applications easily. It has a small and easy-to-extend core. Flask comes under a micro-framework, it has little to no dependencies on external libraries.

AWS: Amazon Web Services (AWS) provides you with trustworthy, scalable, and cost-effective computational resources for any applications. AWS offers a massive global cloud infrastructure that enables us to quickly create, test, and iterate. Rather than waiting weeks or months for hardware, users can rapidly deploy new apps, scale up as their workload expands, and reduce resources as demand decreases. That's why AWS is preferable for web hosting.

METHODOLOGY

A single algorithm may not be able to make the perfect prediction for any given dataset. As a result of their limitations, machine learning algorithms are not able to produce accurate models.

Accuracy counts all of the true predicted values, however not specific for each label that exists. To completely remove this, we use various algorithms.

If we build and combine multiple models, the overall accuracy can be increased. The combination can be implemented by aggregating the outputs from different models with two objectives: reducing the model error and maintaining its generalization.

Some techniques can be used to implement such aggregation.

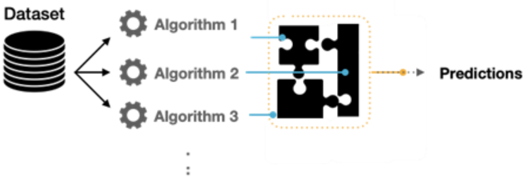


Fig 1: Multiple model combination

The ensemble[5] model construction did not focus solely on the variance of the algorithm used. In this case, each model learns a specific pattern specialized in predicting one aspect, which is known as a weak learner, and those weak learners may use different strategies to map the features with different decision limits.

Bagging: Using bagging, training data is made available for an iterative learning process. Each model learns the error that is produced by the previous model using a slightly different subset of the training datasets. Bagging helps us to reduce the variance and overfitting.

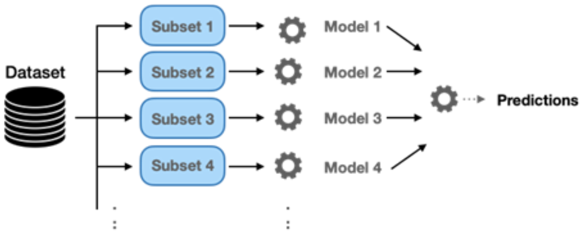


Fig 2: Bagging methodology

Boosting: Adaptive Boosting is an ensemble of algorithms, where we build models on the top of several weak learners. As a result of such adaptability, sequential decision trees adjust their weights based on prior knowledge of accuracy. Hence, we perform the training in such a technique in a sequential rather than parallel process. In this technique, the process of training and measuring the error in estimates can be repeated for a given number of iterations or when the error rate is not changing significantly.

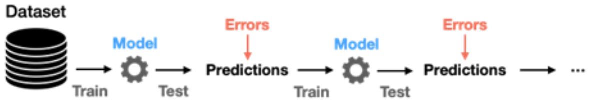


Fig 3: Boosting methodology

Gradient Boosting: With high predictive performance, gradient boosting algorithms are great techniques.

Stacking: Stacking is analogous to boosting models; they produce more robust predictors. Stacking involves building a stronger model from all weak predictions made by weak learners.

Decision tree: A decision tree (DT) is one of the earliest and most prominent machine learning algorithms. A decision tree models the decision logic, i.e., how to classify data items into a tree-like structure based on the results of tests. The nodes of a Decision tree normally have multiple levels where the first or top-most node is known as a root node. All internal nodes represent tests on input variables or attributes. According to the results of the test, the branches of the separation algorithm point to the correct location of the child where the process of testing and integrating reaches the leaf position again. Terminal nodes are the outcomes of the decision process. A common component of medical diagnostic protocols, DTs are easy to interpret and simple to learn.

Random forest: A random forest (RF) is an ensemble classifier[6] consisting of many DTs similar to the way a forest is a collection of many trees. Random subsets are built from the original dataset. To identify the appropriate split, each node in the decision tree considers a random set of features. A decision tree model is fitted to each subset. All decision trees are averaged together to determine the final prediction.

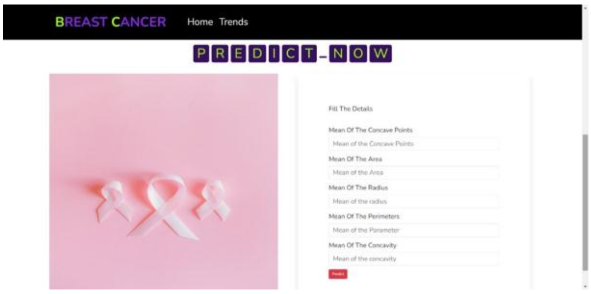
DTs that are grown very deep often cause overfitting of the training data, resulting from a high variation in classification outcome for a small change in the input data. Their training data is quite sensitive, which causes them to make errors on the test dataset. Training datasets use different parts to train the different DTs of an RF. The input vector of a new sample is used to classify it. For a sample to be classified, it must pass down to each DT in a forest. Each DT then The classification is based on a different part of the input vector. The forest then selects the classification based on the most votes (for discrete classification outcomes) or the average of all trees in the forest (for numeric classification outcomes). In addition to considering the results of many different DTs, the RF algorithm can reduce the results of a single DT of the same dataset.

RESULT

We are just beginning to understand what can be done with Big Data in the healthcare industry today. Big Data will be able to generate innovative and novel solutions to the problems of healthcare as a result of the intersection of data from multiple sources, tools, and technologies. There are a lot of implementations that can be achieved using a massive amount of data that every part of healthcare is collecting daily from its customers, types of diseases, and many more.

Likewise, there are a lot of advancements going on in the healthcare industry but there are lots and lots of advancements that need to be introduced in it to let people get interested in their health and their loved ones. And for the same purpose, it has been developed to let people understand themselves better with the help of Machine Learning and Artificial Intelligence which is going to revolutionize this. It is solving many more problems so easily in a fraction of a second with its implementations.

In this, predictions have been made on some diseases for people by taking some values from the user and giving them a proper result based on their inputs. Through this, it is going to get a step closer to predicting diseases for the user. Here are some samples of how it has taken inputs and how you are going to get output based on the information that is provided in the columns.



Fill The Details

Mean Of The Concave Points: Mean of the Concave Points

Mean Of The Area: Mean of the Area

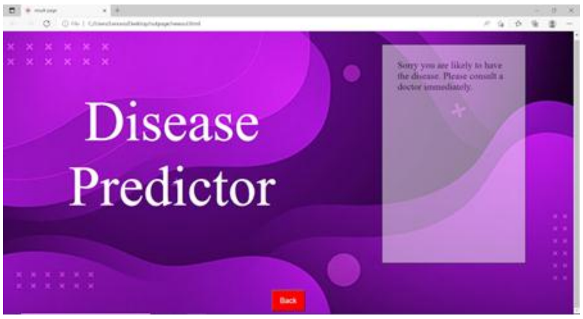
Mean Of The Radius: Mean of the radius

Mean Of The Perimeters: Mean of the Parameter

Mean Of The Concavity: Mean of the concavity

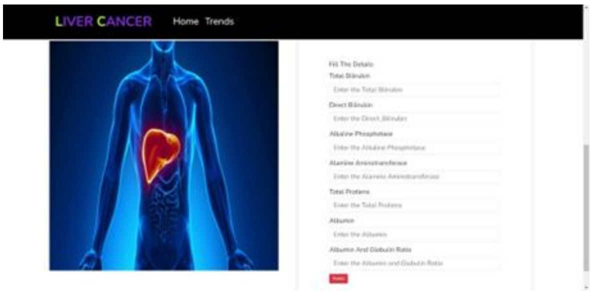
Fig 4: Breast cancer UI

In the above image, as you can see, user can predict their chances of getting breast cancer by providing certain values like the Mean of the concave points, the Mean of the area, the Mean of the radius, etc. By calculating these values they are going to get output telling them their chances of having breast cancer as seen in the below image.



Sorry you are likely to have the disease. Please consult a doctor immediately.

Fig 5: Positive result



Fill The Details

Total Bilirubin: Enter the Total Bilirubin

Direct Bilirubin: Enter the Direct_Bilirubin

Alkaline Phosphotase: Enter the Alkaline Phosphotase

Alamine Aminotransferase: Enter the Alamine Aminotransferase

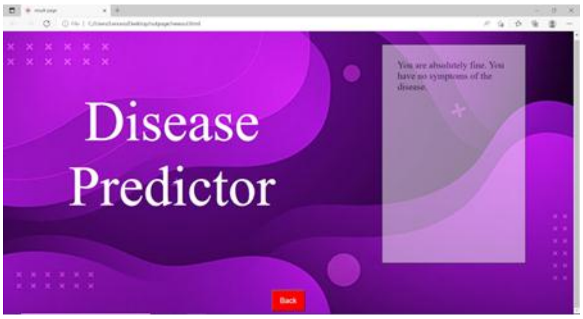
Total Proteins: Enter the Total Proteins

Albumin: Enter the Albumin

Albumin And Globulin Ratio: Enter the Albumin and Globulin Ratio

Fig 6: Lung cancer UI

If someone wants to predict their chances of having lung cancer then, as shown in the above image, they can also predict it by adding certain values asked like Albumin, Albumin and Globulin Ratio, Total Proteins, etc. which will allow them to predict their disease in an early phase. And also the user will get the output based on the information provided as you can see in the below image.



You are absolutely fine. You have no symptoms of the disease.

Fig 7: Negative result

After predicting some of the diseases, we have taken the same diseases to analyze the data of that disease. We have used Tableau and Python to analyze the data which is being fetched from Kaggle itself. Analysis of the data can also be modified by making some adjustments for which we have included some toggle buttons and sliders which will allow the user to adjust the ranges according to which they can know about the data that they want. For their convenience, some labels have been included to guide them to understand the graphs and they will be having certain values based on the ranges and data available at the moment. For example, how many pregnant ladies can have the risk of catching diabetes on an average whether it is Benign or Malignant at a particular age? There is a lot more information that can be consumed from the graphs and people can educate themselves by taking a look at those graphs. Here are some samples of these graphs which a person can see after selecting a particular disease on the webpage.

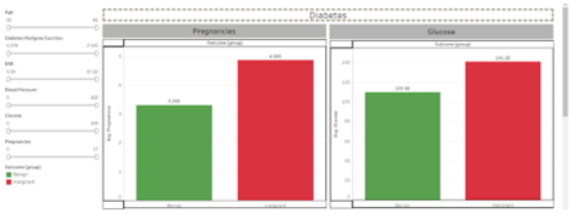


Fig 8: Diabetes analytics UI

In the above image, it is shown that, between the age of 21-81, with a glucose level of 0-199, blood pressure between 0-122, and based on many more factors, how many average pregnancies can catch diabetes, which can be benign(Green bar) or malignant(Red bar) and it is also shown for the glucose levels of diabetic patients at that point.

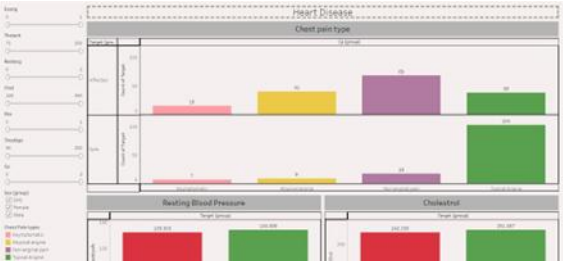


Fig 9: Heart disease analytics UI

In the above image, the graphs are being shown for having heart disease. There are various factors which are affecting for the type of chest pain that a person can have while having heart disease in the above graph it is having Thalach range between 71-202, Cp range of 0-3, Trestbps range of 94-200, Fbs between 0-1 and many more ranges are being used to analyze this data. It is also shown in the graph about who all are safe and who can be affected by giving accurate numbers on an average using different ranges and inputs.

CONCLUSIONS

In today's world, a lot more diseases can affect people in many ways but there are very less technologies that can be accessed and used by the user itself to self-educate or to get some information on this serious topic. Many technological geeks are working in the direction of educating people about the health of themselves and their family members. If a user is educated and has access to the information which is appropriate and precise then it is easy to diagnose the diseases and also people can get alertness or consciousness in themselves which can help them get treated before it gets worse.

So, through this project, which is backed by proper research, we want to get ahead with the help of technology in predicting people's health by using Machine Learning and AI and also analyzing some data for them so that they can gain some knowledge through this website. The sole motive of this project is to give people the freedom to live and take care of their health and we who are here have a little bit of interest in technology and want people to know about what Artificial Intelligence can do for them and how it would be framed for our future selves. In this project, we have used some latest technologies to build a web application that can show your chances of catching some diseases by getting some values from the users. It is a prediction model which can show information about your disease by asking you for some information. And in the second part, we have analyzed some datasets which will be having dynamic values by which one can know about a particular disease, e.g., at what age there are more chances of getting breast cancer or which types of things will let you catch a heart disease. There will be graphs to show this information to the user.

In this, we have leveraged the amount of data that is available out there which can be made of use if someone knows how to use it properly. We have tried to make this information meaningful so that these new technologies can be introduced to the people and can be of use in any way.

REFERENCES

[1] H. Alharthi, "Healthcare predictive analytics: An overview with a focus on Saudi Arabia," Journal of Infection and Public Health, pp. 749–756, Feb. 2018. https://doi.org/10.1016/j.jiph.2018.02.005

[2] Divya Jain, Vijendra Singh, " Feature selection and classification systems for chronic disease prediction: A review", Egyptian Informatics Journal, pp.179-189, Mar. 2018. https://doi.org/10.1016/j.eij.2018.03.002

[3] Atreyi Kankanhalli, Jungpil Hahn, Sharon Ta, Gordon Gao, "Big data and analytics in healthcare: Introduction to the special section", Business media Newyork, Mar. 2016. https://link.springer.com/content/pdf/10.1007/s10796-016-9641-2.pdf

[4] A.Rishika Reddy, P. Suresh Kumar. " Predictive Big data analytics in healthcare", Second International Conference on Computational Intelligence Communicational Technology (CICT), Feb. 2016. https://sci-hub.hkvisa.net/10.1109/CICT.201 6.129

[5] Mohammed Alhamid, "A guide to learning Ensemble technique", Toward Data Science, Mar. 2021. https://towardsdatascience.com/ensemble models-5a62d4f4cb0c

[6] An introduction to Random forest Classifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[7] P. Saranya, Dr. P. Asha, "Survey on Big Data Analytics in Health Care", 2019 International Conference on Smart System and Inventive Technology(ICSSIT), Feb. 2020. https://ieeexplore.ieee.org/document/8987 882

[8] Dhiraj Dahiwade, Gajanan Patle, Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach",2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Aug. 2019. https://ieeexplore.ieee.org/document/8819 782

[9] "Cleveland Heart Disease Dataset," accessed on 03-02-2017.

[10] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," The American journal of cardiology, pp. 304–310, 1989. https://www.ajconline.org/article/0002-9149(89)90524-9/pdf

[11] B. Edmonds, "Using Localised 'gossip' to structure distributed learning", Jan. 2005. https://www.researchgate.net/publication/28764359_Using_Localised_'Gossip'_to_Structure_Distributed_Learning

[12] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, Lin Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities", pp 8869 - 8879, April 2017. https://ieeexplore.ieee.org/abstract/document/7912315