**Implementation of Machine learning algorithms in the field of Bioinformatics**

Mohan Dev Vashisht\*1, Varun Saxena2, Ishita Uniyal3, Dr. Neha Yadav4

Dept. of Artificial Intelligence and Machine Learning

(ADGIPS, Delhi)

mdvashisht29@gmail.com\*

**ABSTRACT**

The amount of data around us in every field is growing exponentially day by day. Our knowledge in the field of Artificial intelligence and Machine learning is also growing as an aftermath of it. One such field that is prevalent in today’s time Is of Bio-informatics with a special emphasis in the field of disease prediction. The study and implementation of various ML algorithms to accomplish this is discussed in this research paper. The development of such model is also discussed using the most efficient pathway to accomplish it.

Keywords: Bioinformatics, Disease Prediction, Feature Scaling, Logistic Regression, SVM

1. **Introduction**

The healthcare sector is considered as the backbone of any country. The quality and the functioning of the healthcare sector is considered as a good factor to measure a country’s growth. India has an advanced healthcare system too with modern cutting-edge technologies. Recent developments in this sector have fueled the development of models for initial detection and later help in making the required drugs that have completely eradicated certain diseases from the entire population.

However, certain diseases still exist in today’s times, for which a cure or a treatment is yet to be established. Diseases like these include Dementia, Cancer, Parkinson's disease etc. [1] According to a recent survey conducted in 2017, the top 15 conditions that accounted for the most DALYs were mostly those causing mortality, such as ischemic heart disease, chronic respiratory diseases, cancer, stroke, and tuberculosis. The Patients suffering from such diseases are left with no other choice, than to just rely on short term treatments for them.

To identify these diseases in early stages and tackle such a condition, an efficient solution can be to develop an algorithm using AI and machine learning which can assist professionals and individuals to develop accurate and efficient cures and treatment processes. [4]

1. **Literature Review**
2. *Related work*

This problem is not only prevalent in India but rather in the entire world. This has fueled a variety of different researches in this field. A few papers have examined the use of (DL) algorithms particularly working on QSAR models that integrate computer and statistical techniques in order to make a theoretical prediction.[7] This section aims to briefly review and summarize these related studies.

Jian Huang et al. (2022) in their study emphasized on the the use of image recognition technology to achieve this task.

Robert Hoehndorf et al. (2021) in their study talked about the predictions of drug targets and their binding affinities using phenotypic effect of drugs.

Woo Suk Tae et al. (2018) in their study emphasied on the clinical role of DTI in various disease processes such as amyotrophic lateral sclerosis, multiple sclerosis, Parkinson's disease, Alzheimer's dementia, epilepsy, ischemic stroke, stroke with motor or language impairment, traumatic brain injury, spinal cord injury, and depression. Valuable DTI preprocessing tools for clinical research are also introduced.[7]

Major success has been achieved through previous ventures but day by day, even more optimized techniques to implement a machine learning algorithm are designed and using these techniques in this field can further optimize the efficiency of the developed systems.

1. *Our contribution*

There are numerous machine learning and deep learning algorithms being used in the field of bioinformatics. The results are highly accurate as well. However, it’s a general trend that the algorithms designed are rather general and not disease specific with a relatively finite number variable. It’s challenging to develop a general algorithm that studies the variables that surround a disease equally and predicts weather the disease is there or not. Our algorithm is disease specific in its nature of study as well as takes more and more variables into account to further increase the accuracy of the prediction. In this project and research, we create a machine learning model based on logistic regression and SVM to achieve the tasks mentioned earlier and also study the implementation of new techniques in the field that are not thoroughly explored yet.[5]

1. **Data**
2. *Dataset*

Fine needle aspiration is a type of biopsy procedure. In fine needle aspiration, a thin needle is

inserted into an area of abnormal-appearing tissue or body fluid. As with other types of

biopsies, the sample collected during fine needle aspiration can help make a diagnosis or rule

out conditions such as cancer. A preexisting labelled dataset consisting of 698 rows and 10 columns are used for this project, it is extracted from a public repository of the University of Wisconsin contributed in the year 1995. The attributes like texture\_mean, area\_mean, smoothness\_mean, symmetry\_mean etc. are stored in a database and this is loaded as the dataset to the algorithm.

1. *Preprocessing*

The data which pertains to healthcare and mainly disease prediction is generally refined and accurate, it consists of data points assembled together based on months and years of surveys and tests. The dataset we used consisted of similar data points i.e. existing symptoms, response to medical tests, response to medical drugs etc.

However, to ensure improved performance and dependability, basic preprocessing was performed on the data. The approaches are summarized below.

***ID Mapping:*** The unique IDs designated to each responding patients were mapped from all the elements of the database and clubbed together before the final study and modelling of the dataset. This ensures that any form of overfitting or underfitting is avoided as the model runs it course and at the same time the accuracy of predictions is preserved.

***Handling missing values: The*** dataset consisted several missing values which were represented by a ‘?’.

If a model detects a ‘?’, it simply identifies it as an object rather than an integer, so that is why we allocated the missing values a numerical value of -99999 which is an integer and at the same time, it is mathematically insignificant to hamper the accuracy of the prediction in any way.

***Use of Lambda function:*** We know, Logistic Regression works on binary datapoints, so by using the lambda function, we designate a binary code of ‘0’ and ‘1’to our compounds in such a way that the algorithm of logistic regression can readily run on the dataset.

***Feature Scaling:*** Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization

1. **Implementation**
2. *Algorithms*

Machine learning (ML) algorithms represent sophisticated computational tools that exhibit adaptability through the analysis of data to anticipate forthcoming outcomes. These algorithms operate on mathematical frameworks, enabling computers to assimilate information, discern patterns, make prognostications, or execute tasks autonomously, without explicit programming instructions. Enclosed within this section are delineations of several algorithms utilized in the development of our model.

***Logistic Regression***: Logistic regression emerges as a supervised machine learning algorithm predominantly employed for binary classification endeavors, where it gauges the probability associated with a specific outcome, event, or observation. This model furnishes a binary or dichotomous output, delineated by two conceivable outcomes, often denoted as yes/no, 0/1, or true/false. Visually, the logistic regression curve exhibits an "S" shape when plotted against the data points [12]

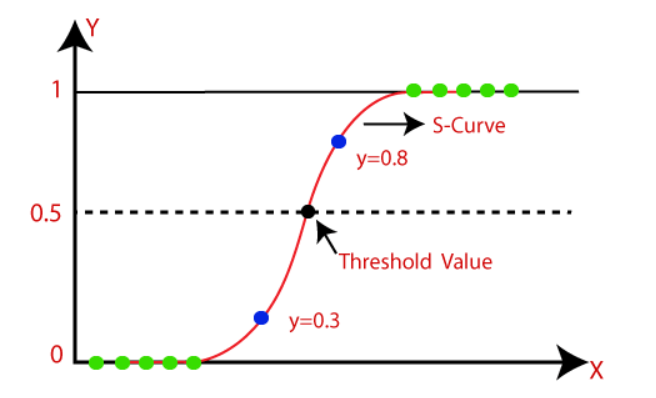


Fig.1: A graphical representation of Logistic Regression

***Support Vector Machine***: Support Vector Machine (SVM) is widely recognized as a top-tier Supervised Learning technique, adept in addressing both Classification and Regression tasks. However, its primary utility predominantly lies in handling Classification challenges within the realm of Machine Learning. At its core, the SVM algorithm aims to craft an optimal decision boundary or line, effectively partitioning n-dimensional space into discernible classes, thereby enabling seamless categorization of future data points into their appropriate categories. This optimal boundary is commonly referred to as a hyperplane.[9]

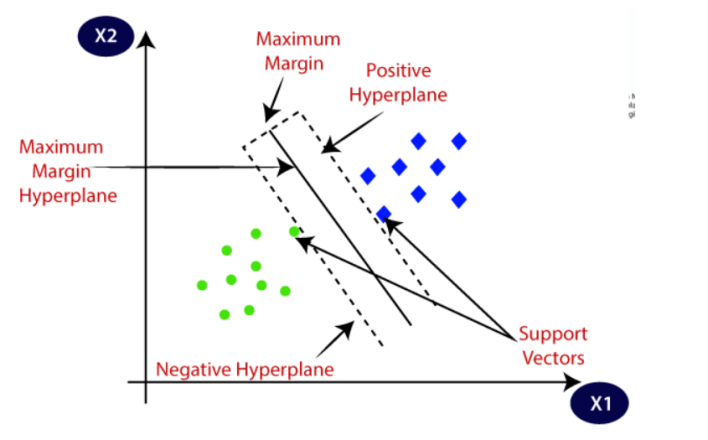


Fig. 2: A graphical representation of Support Vector Machine

1. *Libraries*

***NumPy***: NumPy, or Numerical Python, serves as a Python library that equips users with a range of data structures and mathematical functions tailored for scientific computing. It plays a pivotal role in machine learning by enabling efficient handling and analysis of large dataset.

***Pandas***: Pandas stands as a freely accessible software library, constructed upon NumPy, designed to facilitate data manipulation and analysis. Renowned for its potency in data analysis, Pandas enjoys widespread adoption within the realm of machine learning. It furnishes a rich assortment of data structures and functionalities tailored for handling structured (tabular, multidimensional, potentially heterogeneous) and time series data.

***Scikit-learn***: Scikit-learn represents an open-source library for data analysis, renowned as the pinnacle of Machine Learning (ML) within the Python ecosystem. Prominent aspects and functionalities encompass algorithmic decision-making techniques, encompassing classification tasks, which entail the identification and categorization of data based on perceptible patterns. [18]

***Pickle***: Pickle, a Python module, facilitates the serialization and deserialization of objects into a binary format. Its versatility extends beyond just machine learning models, as it can be applied to serialize and deserialize any object as needed.

1. **Methodology**
2. *Type of SVC used*

The type of SVC used to design the algorithm is Linear Support Vector Machine (Linear SVC) is an algorithm that attempts to find a hyperplane to maximize the distance between classified samples[19]

1. Type of Logistic Regression used

The type of Logistic Regression used is liblinear using

sklearn.linear\_model.LogisticRegression

Liblinear is a solver in scikit-learn that supports logistic regression (LR) and linear support vector machines. It uses a Coordinate Descent (CD) algorithm to solve optimization problems by performing approximate minimization along coordinate directions or coordinate hyperplanes.[20]

1. **Bifurcation in predictions**

Broadly speaking, there are 2main types of tumours namely “Benign” and “Malignant”.

***Benign***: Benign tumors are non-cancerous and generally less harmful. Benign tumors have distinct, smooth, regular borders, grow slowly, and remain in their primary location.

***Malignant***: A malignant tumor is cancerous and can spread to other parts of the body through the

lymphatic system or bloodstream, a process known as metastasis. It is non-capsulated with poorly defined shapes. They grow rapidly and can have high levels of kidney damage.

1. **Results**

Through a well-defined approach, we developed a machine learning model using the above mentioned algorithms and libraries which was able to predict weather a patient is suffering from breast cancer or not. It was an example of binary prediction where Maligant represents presence of breast cancer whereas Benign represents the non-existence of it.

A snapshot of the prediction given by our model is attached below-

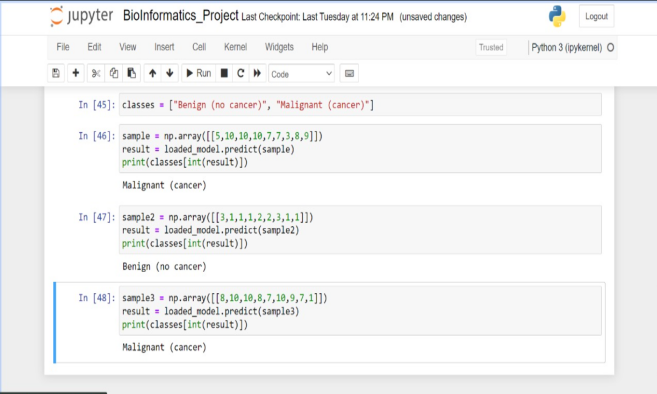


Fig.3: A snapshot of the predictions carried out by the model

1. **Conclusion & Scope for Future Work**

As our population increases day by day, the need for a healthcare system which constantly evolves over time and adapts to changing circumstances have become crucial for the well-being of any nation and its people. According to reports, only 13% of rural population in India have access to primary health centers, only 43.5% of children in India receive all vaccinations against harmful diseases and these statistics go on. The need to work on improving these numbers is of the highest priority in the times that follow. Thankfully more and more research and study is being done in this area and new ways of incorporating automation in this sector is the talk of the hour. We have models working on curing these types of diseases which are being constantly optimized using the new technologies around us. Logistically, we need to ensure that these models are deployed on relevant platforms so that they reach to the rural population as well through an organized pathway. Government support for research in this area can also lead to much more positive results for us. Lastly, there is scope for future work in this sector as well, with the introduction and increasing popularity of Deep Neural Networks, we can now develop models that can function exactly like a human brain and health specialists can work with those models to further optimize their working procedures. Use of Generative Adversial Networks have been fully implemented in the medical field and will be more widely used in clinical medicine in the coming times. Automation and AI will definitely rule the future.

REFERENCES

[1] Arthur M Lesk, “Bioinformatics: Genomics, Proteomics & Data Analysis”. USA:2002

[2] Bioinformatics : A way to decipher DNA and cure life’s deadlist diseases | Spencer Hall | TEDxUGA

[3] Noam Auslander, Ayal B. Gussow and Eugene V. Koonin : “Incorporating Machine Learning into Established Bioinformatics Frameworks” , MDPI, USA (2021)

[4] Nareen O. M. Salim, Adnan Mohsin Abdulazeez: “Human Diseases Detection Based On Machine Learning Algorithms”: A Review in proceedings of the IJSAB, Iraq (2021)

[5] Samin Poudel : “A Study of Disease Diagnosis Using Machine Learning” ,MDPI, USA (2022)

[6] Heba Askr et al. : Importance of latest technology in DL application for the DTIs such as XAI and DTs in Springer, Egypt (2022)

[7] Heba Askr, Enas Elgeldawi,Heba Aboul Ella,Yaseen A. M. M. Elshaier , Mamdouh M. Gomaa, Aboul Ella Hassan : “Deep learning in drug discovery: an integrative review and future challenges” in proceedings of the NIH, Egypt (2022)

[8] Robert Hoehndorf et al.: “The predictions of drug targets and their binding affinities using phenotypic effect of drugs” in proceedings of the ISCB, Saudi Arabia (2021)

[9] Woo Suk Tae et al. “The clinical role of DTI in various disease processes”, NIH, South Korea (2018)

[10] Richard C. Mohs, Nigel H. Greig: “Drug discovery and development: Role of basic biological research”, NIH, USA (2017)

[11] Joanne Peng, Kuk Lida Lee: “An Introduction to Logistic Regression Analysis and Reporting, Journal of Educational Research”, USA (2002)

[12] Pal, Ankita Logistic Regression: A simple primer, India (2021)

[13] Ashok Panagariya: “The Challenges and innovative solutions to rural health dilemma”, NIH, USA (2014)

[14] Arup Jana: “Prevalence and potential determinants of chronic disease among elderly in India: Rural-urban perspectives”, NIH, USA (2022)

[15] Fan Zhang, Luyao Wang, Jiayin Zhao, Xinhong Zhang: “Medical applications of generative adversarial network a visualization analysis” , NIH, USA (2023)

[16] National Library of Medicine

Available at : <https://www.ncbi.nlm.nih.gov/>

Last accessed : 10/04/2024

[17] ResearchGate

Available at : <https://www.researchgate.net/>

Last accessed : 15/04/2024

1. The Apache Software Foundation

Last accessed : 7/04/2024

[19] Scikit-learn

Available at : <https://scikit-learn.org/stable/index.html>

Last accessed : 2/04/2024