

# Implementation of Machine Learning Algorithms in the Field of Bioinformatics

Mohan Dev Vashisht<sup>\*1</sup>, Varun Saxena<sup>1</sup>, Ishita Uniyal<sup>1</sup>, Neha Yadav<sup>1</sup>, Mohd Izhar<sup>2</sup>

<sup>1</sup>Dept. of Artificial Intelligence and Machine Learning, Dr. Akhilesh Das Gupta Institute of Professional Studies, New Delhi, 110053, Delhi, India

<sup>2</sup>ADGIPS, FC-26, Panduk Shila Marg, Zero Pusta Rd, Shastri Park, Shahdara, New Delhi, Delhi 110053

\* Corresponding author

doi: <https://doi.org/10.21467/proceedings.178.13>

## ABSTRACT

With the rapid growth of data across various fields, advancements in Artificial Intelligence (AI) and Machine Learning (ML) have gained significant momentum. One area where this progress has had a profound impact is bioinformatics, particularly in disease prediction. The availability of vast amounts of biological data has opened the door to leveraging ML algorithms to identify patterns and make predictions that could transform healthcare. This paper focuses on the application of ML in bioinformatics, with a special emphasis on disease prediction. It explores the use of various ML algorithms to achieve accurate and meaningful results. By harnessing these tools, researchers can analyze complex datasets more efficiently and uncover insights that were previously difficult to detect. The study also discusses the process of developing predictive models, highlighting methods that ensure efficiency and reliability. By addressing challenges and presenting solutions, the research illustrates how ML can be applied to tackle critical healthcare issues. This work emphasizes the potential of combining computational techniques with biological research to advance disease prediction and improve diagnostics, paving the way for more personalized treatment approaches and better patient outcomes.

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

## I. 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. Identifying these diseases in early stages is a crucial step in tackling such conditions, 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 [2]. 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



© 2025 Copyright held by the author(s). Published by AIJR Publisher in "Proceedings of 2<sup>nd</sup> International Conference on Emerging Applications of Artificial Intelligence, Machine Learning and Cybersecurity" (ICAMC 2024). Organized by HMR Institute of Technology and Management, New Delhi, India on 16-17 May 2024.

Proceedings DOI: [10.21467/proceedings.178](https://doi.org/10.21467/proceedings.178); Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-984081-8-1

predicts whether the disease is there or not. The algorithm in this study 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 study, a machine learning model is created using 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.

## II. Literature Review

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 [3]. This section aims to briefly review and summarize these related studies.

[4] in their study, the authors emphasized on the use of image recognition technology to achieve this task. [5] in their study, the authors talked about the predictions of drug targets and their binding affinities using phenotypic effect of drugs. [6] their study emphasized 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. 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.

## III. Data

### A. 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 the mean, area, smoothness and symmetry of the tumor are stored in a database and this is loaded as the dataset to the algorithm.

### B. 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. In the Identity mapping process, 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 its course and at the same time the accuracy of predictions is preserved. 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. 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 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

## IV. Implementation

### A. 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, the binary class is allotted on the basis of whether the threshold value is exceeded or not. [7]

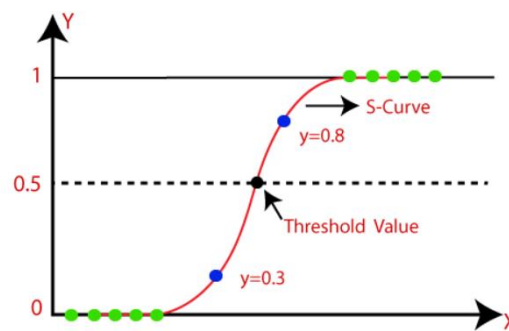


Figure 1: A graphical representation of Logistic Regression

**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.

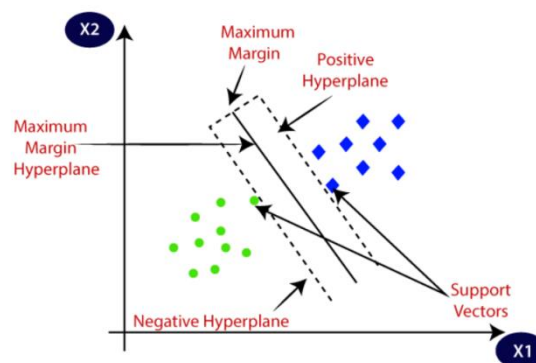


Figure 2: A graphical representation of Support Vector Machine

## B. 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. [8]

**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.

## V. Methodology

To develop the breast cancer prediction model, two machine learning algorithms were employed: Linear Support Vector Machine (Linear SVC) and Logistic Regression. These algorithms were chosen for their ability to effectively classify binary outcomes such as benign or malignant tumors. The Linear SVC algorithm was utilized to design a hyperplane that maximizes the margin between classified samples. This technique is particularly effective for binary classification tasks, as it aims to minimize misclassification by maximizing the distance between data points of different classes. The model's performance was evaluated based on its ability to correctly separate benign and malignant tumor samples. The Logistic Regression model was implemented using the liblinear solver from the sklearn.linear\_model.LogisticRegression library. The liblinear solver is well-suited for smaller datasets and binary classification problems. It employs the Coordinate Descent (CD) algorithm, which optimizes the model by iteratively performing approximate minimization along individual coordinate directions or coordinate hyperplanes. This approach ensures efficient convergence to an optimal solution. Both models were trained and tested on the dataset, and their performances were evaluated using metrics such as accuracy, precision, recall, and F1-score. These metrics provided a comprehensive understanding of each model's effectiveness in predicting breast cancer.

## VI. Bifurcation in predictions

Broadly speaking, there are two main types of tumors namely “Benign” and “Malignant”.

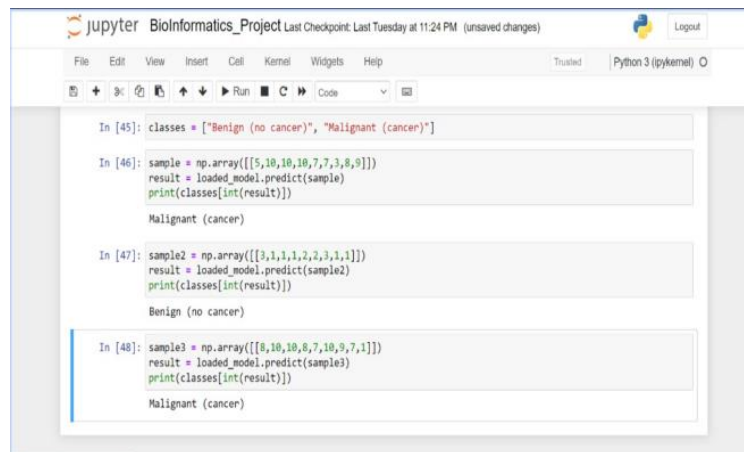
**Benign:** Benign tumors are non-cancerous and generally less harmful. Benign tumors have distinct, smooth, regular borders, grow slowly, and remain in their primary location.

**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 exists with poorly defined shapes. They grow rapidly and can have high levels of kidney damage.

## VII. Results

In our study, a machine learning model was developed to predict whether a patient is affected by breast cancer. This binary classification problem involved distinguishing between two categories: *Malignant* and *Benign*. The model was built using a well-defined pipeline, leveraging algorithms and libraries that

facilitated accurate predictions. After training the model on a comprehensive dataset, the predictive capability was validated through test samples. The overall accuracy of the model was recorded at 93%, demonstrating its reliability in classification tasks. The snapshot of the result in our Jupyter Notebook highlights specific predictions made by the model. We tested the model with various samples, each consisting of critical input features such as texture, thickness, smoothness, and other relevant attributes. These predictions showcase the model's ability to accurately classify the input data based on learned patterns. The predictions align with the ground truth, reflecting the robustness of the algorithm. By handling missing data during preprocessing and focusing on the significant features, we ensured that the model remained effective without overfitting. The results underscore the potential of machine learning in automating breast cancer diagnosis, especially in scenarios where immediate medical evaluation may not be available. This model serves as a promising step toward developing accessible diagnostic tools for improving healthcare outcomes. A snippet of the prediction given by our model is attached below-



```

In [45]: classes = ["Benign (no cancer)", "Malignant (cancer)"]

In [46]: sample = np.array([[5,10,10,10,7,7,3,8,9]])
        result = loaded_model.predict(sample)
        print(classes[int(result)])
        Malignant (cancer)

In [47]: sample2 = np.array([[9,1,1,1,2,2,3,1,1]])
        result = loaded_model.predict(sample2)
        print(classes[int(result)])
        Benign (no cancer)

In [48]: sample3 = np.array([[8,10,10,8,7,10,9,7,1]])
        result = loaded_model.predict(sample3)
        print(classes[int(result)])
        Malignant (cancer)

```

Figure 3: A snippet of the predictions carried out by the model

### VIII. 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 constant development 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 studies are being done in this area and new ways of incorporating automation in this sector is the talk of the hour. Currently, the healthcare sector has various models working on curing these types of diseases which are being constantly optimized using the new age 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 Adversarial 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.

## Declarations

## Competing Interests

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## How to Cite

Mohan Dev Vashisht, Varun Saxena, Ishita Uniyal, Neha Yadav, Mohd Izhar (2025). Implementation of Machine Learning Algorithms in the Field of Bioinformatics. *AIJR Proceedings*, 111-116. <https://doi.org/10.21467/proceedings.178.13>

## REFERENCES

- [1] A. M. Lesk, *Bioinformatics: Genomics, Proteomics & Data Analysis*. USA: Academia.,2002
- [2] S. Hall, "Bioinformatics: A way to decipher DNA and cure life's deadliest diseases," presented at TEDxUGA, [Online]. Available: <https://www.ted.com/tedx>.
- [3] H. Askr, E. Elgeldawi, H. Aboul Ella, *et al.*, "Deep learning in drug discovery: an integrative review and future challenges," *Artificial Intelligence Review*, vol. 56, pp. 5975–6037, 2023. [Online]. Available: <https://doi.org/10.1007/s10462-022-10306-1>
- [4] J. Huang, J. Li, Z. Li, Z. Zhu, C. Shen, G. Qi, and G. Yu, "Detection of diseases using machine learning image recognition technology in artificial intelligence," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–14, 2022. [Online]. Available: <https://doi.org/10.1155/2022/5658641>
- [5] T. Hinnerichs and R. Hoehndorf, "DTI-Voodoo: machine learning over interaction networks and ontology-based background knowledge predicts drug-target interactions," *Bioinformatics*, vol. 37, no. 24, pp. 4835–4843, Dec. 2021. doi: 10.1093/bioinformatics/btab548. PMID: 34320178; PMCID: PMC8665763.
- [6] W. S. Tae, B. J. Ham, S. B. Pyun, S. H. Kang, and B. J. Kim, "Current clinical applications of diffusion-tensor imaging in neurological disorders," *Journal of Clinical Neurology*, vol. 14, no. 2, pp. 129–140, Apr. 2018. doi: 10.3988/jcn.2018.14.2.129. PMID: 29504292; PMCID: PMC5897194.
- [7A. Pal, "Logistic regression: A simple primer," *Cancer Research, Statistics, and Treatment*, vol. 4, no. 3, pp. 551–554, Jul.–Sep. 2021. doi: 10.4103/crst.crst\_164\_21.
- [8] The Apache Software Foundation. [Online]. Available: <https://www.apache.org>. [Accessed: Jul. 4, 2024].
- [9] Scikit-learn. [Online]. Available: <https://scikit-learn.org/stable/index.html>. [Accessed: Apr. 2, 2024].
- [10] A. Jana and A. Chattopadhyay, "Prevalence and potential determinants of chronic disease among elderly in India: Rural-urban perspectives," *PLoS One*, vol. 17, no. 3, pp. e0264937, Mar. 2022. doi: 10.1371/journal.pone.0264937. PMID: 35275937; PMCID: PMC8916671.
- [11] N. O. M. Salim and A. M. Abdulazeez, "Human diseases detection based on machine learning algorithms: A review," *International Journal of Science and Business*, vol. 5, no. 2, pp. 102–113, 2021. doi: <https://doi.org/10.5281/zenodo.4462858>.