Parichay: Maharaja Surajmal Institute Journal of Applied Research

# Comparative Analysis of Machine Learning Algorithm for Breast Cancer Classification: A Comprehensive Evaluation

**Menal Dahiya*, Zainab Asif and Soumya***


## ABSTRACT

***Breast cancer remains one of the most common and dangerous sickness influencing women (as well as men, to some extent) around the world, with its frequency consistently expanding throughout the long term. Early recognition and precise forecast of breast cancer plays a crucial part in working on results and endurance rates. However, notwithstanding headways in clinical innovation and screening programs, many cases are as yet analysed at cutting edge stages, restricting treatment choices and lessening endurance rates. Because of this basic medical services challenge, this research paper presents an original web application intended for breast cancer prediction, meaning to upgrade early discovery endeavours and work on persistent guess. The web application uses machine learning methods to investigate an exhaustive arrangement of elements removed from Fine Needle Aspiration (FNA) reports, including cell qualities, growth size, and morphology. The integration of Fine Needle Aspiration (FNA) and Machine Learning (ML) offers a logical and innovative solution to the limitations of standalone diagnostic methods [1]. By utilising the Wisconsin Breast Cancer Dataset, a widely recognized repository of breast cancer data, the application is trained to classify breast cancer cases into two categories: malignant and benign.***


***Keywords: Breast Cancer Prediction, Machine Learning (ML) models, Model Comparison, Classifier, Training, Accuracy, Precision, Support Vector Machine***

## 1. INTRODUCTION

The most prevalent medical risk encountered by middle-aged women is breast cancer. Enhancing the likelihood of survival from breast cancer hinges on early detection. A cancer prognosis typically involves multiple physicians from different specialties using different subsets of biomarkers and multiple clinical factors, including the age and general health of the patient, the location and type of cancer, as well as the grade and size of the tumor [2]. With the help of latest, efficient and advanced screening methods, the majority of such cancers are diagnosed when the disease is still at a localized stage [3]. Machine Learning, a field of artificial intelligence, has demonstrated remarkable potential in medical diagnostics [4]. The integration of machine learning methodologies in healthcare analytics is steadily gaining momentum. Undoubtedly, the evaluation of patient clinical records and medical professionals' expertise remains paramount in diagnosis.

Employing classification systems can mitigate many potential medical errors and facilitate a more thorough analysis of healthcare data in less time. Precise and timely breast cancer prediction empowers physicians and healthcare providers to make informed decisions regarding patient treatment plans. Breast cancer, a prevalent global malignancy [5], significantly impacts individuals and healthcare systems, underscoring the need for innovative early detection and treatment strategies. The application is devoted to the early detection of bosom disease, aka, breast cancer through the investigation of Fine Needle Aspiration (FNA) reports. Fine Needle Aspiration is a negligibly obtrusive methodology regularly utilized for getting tissue or liquid examples from dubious masses or knots in the breast. This web application uses machine learning and data analysis methods to decipher FNA reports, assisting in the prompt identification of cancerous cells.

The goal of Cancer Guardian, our online application, is to precisely assess and categorize breast cancer data derived from fine needle aspiration, offering a dependable resource for medical professionals. Through the utilization of sophisticated algorithms, our platform seeks to efficiently evaluate this data, distinguishing between benign and malignant cases, thereby facilitating prompt and accurate diagnoses. Our objective is to enhance clinical decision-making, elevate patient outcomes, and propel progress in breast cancer diagnosis and therapy. Prospective upgrades may involve integrating with electronic health records systems and refining predictive models continuously to enhance precision and user-friendliness. By garnering widespread acceptance and ongoing enhancements, this web application has the potential to transform breast cancer risk evaluation, ultimately leading to better patient outcomes and public health advancements.

The categorization of cancer cells into benign and malignant necessitated a thorough analysis of the data sourced from UCI. In subsequent sections of the paper, we elucidate the diverse

**Faculty & Students, Department of Computer Applications, Maharaja Surajmal Institute*
*menaldahiya@msijanakpuri.com, zasif1106@gmail.com, soumyashubham1@gmail.com*

methodologies employed in data analysis, encompassing the feeding of data into various models for classification, optimizing these models, attaining requisite outcomes, and integrating these results into the operational framework of the BCP system. Machine learning presents a versatile array of models for addressing classification tasks pertinent to breast cancer diagnosis. Among these, logistic regression stands as a foundational approach, predicting the probability of a binary outcome based on input features. Decision trees delineate the feature space into distinct regions, making decisions predicated on feature values at each node. Support Vector Machines (SVMs) endeavor to ascertain the hyperplane that most effectively segregates classes, maximizing the margin between them. Random Forests, functioning as an ensemble technique, amalgamate multiple decision trees to enhance generalization and resilience. K-Nearest Neighbours (KNN) classify samples by discerning the majority class among their closest neighbors in feature space. Naive Bayes, grounded in Bayes' theorem with strong independence assumptions between features, offers computational efficiency and often exhibits remarkable efficacy, especially with limited datasets.

## 2. RELATED WORKS

Extensive research efforts have been dedicated to leveraging computer algorithms in the diagnosis of breast cancer. Some researchers, like Polat et al., used a method called LS-SVM and got an accuracy of about 98.5% [6]. Akay tried a different method called support vector classification and got around 99% accuracy without using cross-validation [7]. Yeh et al. used statistics and optimization techniques together and achieved about 98.7% accuracy [8]. Marcano-Cedeño et al. used Artificial Neural Networks and reached an accuracy of 99.3% [9]. Another study by Kaya and Uyar focused on detecting hepatitis using a mix of algorithms, and they got an accuracy of about 98.6% [10]. These are just a few examples of how machine learning and data analysis are being used in healthcare to predict and recognize diseases.

Numerous projects related to breast cancer prediction are available on various platforms such as YouTube, GitHub, and other websites. These projects provide valuable resources for exploring algorithms and prerequisites necessary for predicting breast cancer using machine learning techniques. They offer opportunities to learn about the application of machine learning algorithms for classifying and defining breast cancer.

**TABLE 1: list of related authors with references, method technology, and accuracies achieved**

| AUTHOR | METHOD TECHNOLOGY | ACCURACY |
|---|---|---|
| Polat et al | LS-SVM | 98.5% |
| Akay | Support vector classification | 99% |
| Yeh et al. | Statistics and optimization techniques | 98.7% |
| Marcano-CedeÑo et al | Artificial Neural Networks | 99.3% |
| Kaya and Uyar | Mix of algorithms of classifications | 98.6% |

## 3. PROPOSED FRAMEWORK

The problem introduced in the initial section suggests a plan to develop a classification model with enhanced accuracy for predicting breast cancer patients. The framework consists of several key phases:

1. **Selecting the Dataset:** Choosing the appropriate dataset for analysis.
2. **Preprocessing of Data:** Preparing and cleaning the selected data for analysis.
3. **Classifier Training:** Utilizing various algorithms such as Support Vector Machines (SVM), Linear Regression, and K-Nearest Neighbours (KNN) to train the model.
4. **Optimizing the training model:** Refining the trained model to achieve the highest possible accuracy.
5. **Utilising the model for predictions:** Employing the trained model to make predictions.

Each phase involves specific tasks and procedures to effectively build and utilize the classification model for breast cancer prediction.

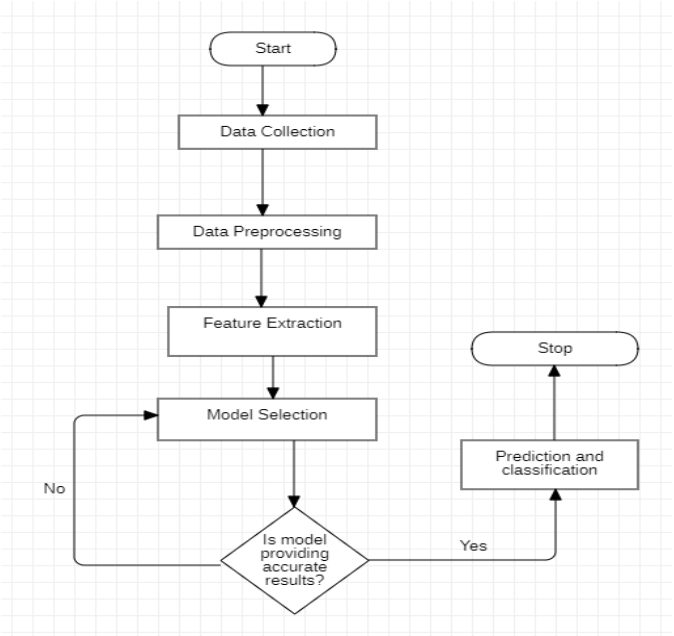


**Fig 1: Workflow of the model**

### *3.1 Selecting the Dataset*

The dataset hails from the UCI repository [16], a renowned source for benchmark datasets. The chosen dataset is the Breast Cancer Wisconsin (Original) dataset, consisting of 699 instances. Within this dataset, 16 instances unfortunately contain missing value. In terms of distribution, about 65.0% of the samples are benign, while the remaining 35.0% are malignant [16]. Also, the UCI repository collected this data based on FNA reports, which is fine needle analysis reports on of men and women, hence, the data proves out to be adequate

for our use. We have taken the Kaggle dataset licensed by UCI, hence 569 entries of male and female having the following distribution, is present in the dataset [17] :

All 32 features have been thoroughly considered in our analysis. However, for more basic classification tasks, one may opt to focus on only 4 to 6 features. Our aim is to attain precise approximations that fulfill the medical objective of determining breast cancer type, thereby obviating the necessity for additional medical procedures in cases of non-cancerous types. After, printing various statistical measures of the data to analyze it, following are the findings:

| | mean radius | mean texture | mean perimeter | mean area | mean smoothness | mean compactness | mean concavity | mean concave points | mean symmetry | mean fractal dimension |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 | 569.000000 |
| mean | 14.127292 | 19.289649 | 91.969033 | 654.889104 | 0.096360 | 0.104341 | 0.088799 | 0.048919 | 0.181162 | 0.062798 |
| std | 3.524049 | 4.301036 | 24.298981 | 351.914129 | 0.014064 | 0.052813 | 0.079720 | 0.038803 | 0.027414 | 0.007060 |
| min | 6.981000 | 9.710000 | 43.790000 | 143.500000 | 0.052630 | 0.019380 | 0.000000 | 0.000000 | 0.106000 | 0.049960 |
| 25% | 11.700000 | 16.170000 | 75.170000 | 420.300000 | 0.086370 | 0.064920 | 0.029560 | 0.020310 | 0.161900 | 0.057700 |
| 50% | 13.370000 | 18.840000 | 86.240000 | 551.100000 | 0.095870 | 0.092630 | 0.061540 | 0.033500 | 0.179200 | 0.061540 |
| 75% | 15.780000 | 21.800000 | 104.100000 | 782.700000 | 0.105300 | 0.130400 | 0.130700 | 0.074000 | 0.195700 | 0.066120 |
| max | 28.110000 | 39.280000 | 188.500000 | 2501.000000 | 0.163400 | 0.345400 | 0.426800 | 0.201200 | 0.304000 | 0.097440 |

**Table 2: Statistical measures and their values depicted from the features**

### *3.2 Preprocessing of Data*

In the Wisconsin dataset there were around 3-4 outliers and missing data. After adjusting the data, we have selected all 32 features for an apt classification (relationship or strongness of dependence between all features). Following is the correlation amongst all of those:

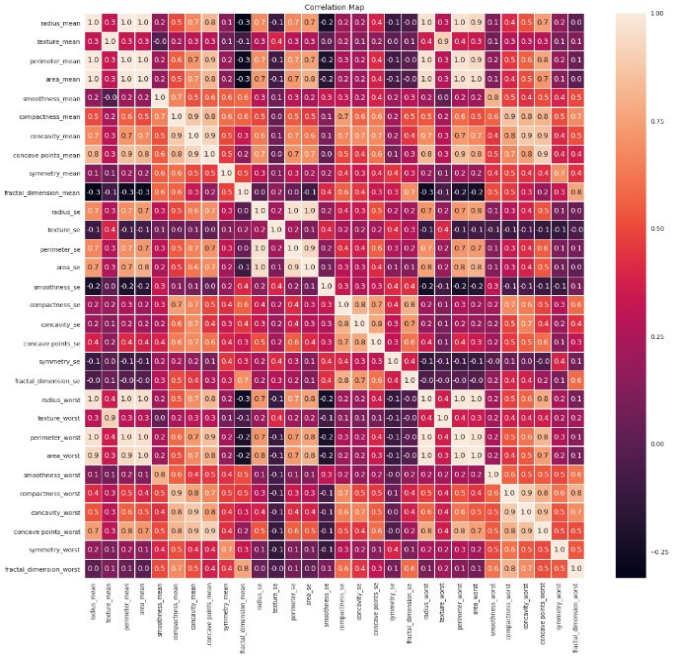


**Fig. 2: Correlation map, showing the dependency of all the features on each other**

### *3.3 Classifier Training*

The classification of these datasets hinges on the identification of distinct attributes exhibited by the sample variables, facilitating their classification into either malignant or benign classes. This methodology entails harnessing learned patterns from training data to forecast outcomes for new, unseen data. Initially, algorithms undergo training on labeled data, utilizing this acquired knowledge to effectively classify unknown samples thereafter. In the context of this study, the objective is to refine accuracy by employing LR, SVM, and KNN classifiers. A comparative study of different kernel functions for breast cancer detection using SVM with different kernel functions using neural network based method using MLP and the affect of selecting feature subsets before applying classification with different kernels is examined [11].The ultimate goal is to ascertain which classifier is most suitable for effectively classifying diabetes. A somewhat newer machine learning technique is called a support vector machine or SVM[12]. Support Vector Machines (SVMs) are a popular machine learning method for classification, regression, and other learning tasks [7]. We have trained or model on support vector machine, the paper specifies further why we have chosen SVM over other algorithms. SVM can be extended for multiclass problems using the so-called one-vs-rest approach [13].

| ALGORITHM | TYPE OF LEARNING | DATA PROCESSING METHOD | ACCURACY | PRECISION | METHOD OF EVALUATION |
|---|---|---|---|---|---|
| SUPPORT VECTOR MACHINES (SVM'S) | SUPERVISED | FEATURE SELECTION | 97.3% | 96% | FUNCTION GRAPH (LR) |
| NAÏVE BAYES | SUPERVISED | FEATURE SELECTION | 93.6% | 94% | CONFUSION MATRIX |
| DECISION TREES | SUPERVISED | FEATURE SELECTION | 95% | 95.52% | CONFUSION MATRIX |
| KNN | UNSUPERVISED | SELECTION OF COMMON FEATURES | 94% | 94% | PERFORMANCE MATRIX |
| RANDOM FOREST | SUPERVISED | FEATURE SELECTION | 97% | 95.6% | BINARY CLASSIFICATION |
| LOGISTIC REGRESSION | SUPERVISED | FEATURE SELECTION | 96.5% | 95% | ACCURACY METHOD |

**Table 2: Comparison report of algorithms used for breast cancer prediction in our web app**

On the Wisconsin Breast Cancer dataset, we evaluated the performance of various models using metrics such as Accuracy and Precision. Our analysis focused on comparing the models based on their predictive capability.

For each model, we generated a confusion matrix, which includes actual and predicted labels, as well as metrics such as True Negative (TN), False Negative (FN), True Positive (TP), and False Positive (FP). These metrics are essential for understanding the performance of the models in correctly classifying breast cancer cases.

CLASSIFICATION USING ACCURACY FORMULA:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100$$

Precision measures the proportion of positive class predictions that are actually positive, providing insights into the model's ability to minimize false positive predictions. Accuracy, calculated using the confusion matrix, indicates the proportion of correctly classified tuples in the training and testing datasets. It provides a general measure of the model's overall performance in correctly classifying instances.

CLASSIFICATION USING PRECISION FORMULA:

$$\text{Precision} = \frac{TP}{TP + FP}$$

By analysing these metrics across different models, we gain insights into their relative strengths and weaknesses in breast cancer prediction. This comprehensive evaluation helps us identify the most effective models for our dataset and guides further refinement and optimization efforts.

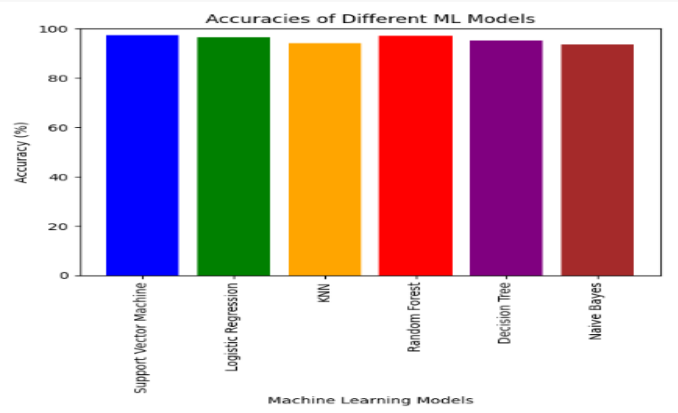


**Fig. 3: Accuracies of different algorithms for the BCP model**

### *3.4 Optimizing the training model*

Optimizing a trained model, Support Vector Machine (SVM) in taking for example, is essential to maximize its predictive accuracy in breast cancer diagnosis. This process involves fine-tuning the model's parameters and optimizing its hyperparameters to enhance its performance. In the case of SVM, key parameters like the choice of kernel function, regularization parameter (C), and kernel coefficient (gamma) play crucial roles in determining the model's effectiveness. Techniques such as grid search or randomized search can be employed to systematically explore the hyperparameter space and identify the combination that yields the best results. Moreover, algorithms like random forest, decision trees and other such algorithms also prove out to be viable in giving good accuracy rates, however , they overfit the date wherein these models give desired outputs with the training set but when fed new data it is not able to recognise defeating the purpose of classification for the type of cancer.

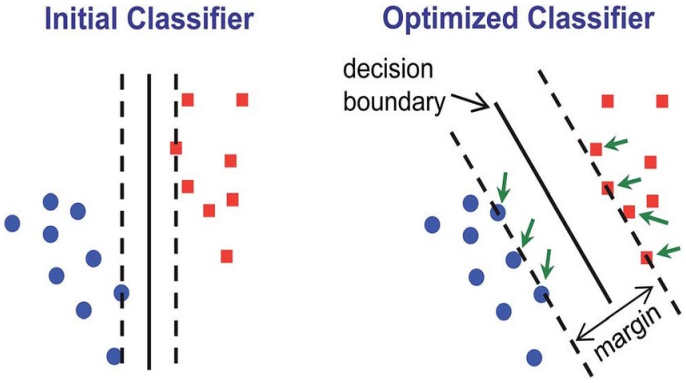


**Fig. 4: Aftereffects of optimizing Support Vector Machine (SVM) classifier**

### *3.5 Utilising the model for predictions*

Using the trained Support Vector Machine (SVM) model to predict breast cancer involves applying it to accurately predict outcomes for new data that it hasn't seen before. Once the SVM model has been trained on a dataset with labelled information relevant to breast cancer diagnosis, it learns patterns and connections within the data. When given new cases, the trained SVM model uses these learned patterns to decide if they belong to the benign or malignant category. This prediction process entails feeding the features of the new cases through the trained SVM model, which then uses its learned parameters to create a decision boundary and classify the cases accordingly. By effectively employing the trained SVM model for prediction, we can provide clinicians with valuable assistance in diagnosing breast cancer, helping with early detection and planning treatment.

## 4. RESULT ANALYSIS

Utilising a trained Support Vector Machine (SVM) model to predict breast cancer involves relying on its ability to accurately predict outcomes for new data. After teaching the SVM with labelled data containing important features for breast cancer diagnosis, it becomes adept at recognizing patterns within the data. When presented with new cases, the SVM uses these learned patterns to determine whether they're benign or malignant. It accomplishes this by analysing the features of the new cases through its trained system, establishing boundaries between different classes based on its learning, and then putting into categories these cases accordingly. This would provide doctors with valuable support in diagnosing breast cancer, facilitating early detection and effective treatment planning.

SVM stands out from other prediction methods because it is quite accurate, holding a success rate of 97.3%, and it is

considerable not overfitting the date (a common problem with other methods like random forest and decision trees). Unlike those, SVM finds an intersection between being complex enough to work well and being general enough to handle new cases smoothly. It does this achievably by drawing lines between different types of cases, making sure it doesn't blur the boundaries between them. In short, using SVM for breast cancer prediction shows useful results, helping doctors/pathologists make better decisions and ultimately improving patient outcomes.

## 5. IMPLEMENTATION

Implementing the trained Support Vector Machine (SVM) model for breast cancer prediction involves several steps, starting with pre-processing new data to ensure it's in the same format as the training data. This may include scaling features, handling missing values, and encoding categorical variables if necessary. Once the data is pre-processed, it can be passed through the trained SVM model to make predictions. Furthermore, it's essential to analyse any misclassifications made by the model to identify patterns or common characteristics among the misclassified instances. This analysis can help in refining the model further or uncovering insights that may be useful for improving the diagnostic process.

Once the SVM model is ready to give the predictions, it can be implemented in the final Web App to predict the breast cancer with the help of the sample which will be gathered by Fine Needle Aspiration (FNA). Overall, the implementation and result analysis phase is crucial for assessing the effectiveness of the trained SVM model in breast cancer prediction, identifying areas for improvement, and gaining insights that can inform future research or clinical practice.


**Fig. 5: The web app, leveraging SVM model and FNA to predict and classify breast cancers**

## 6. CONCLUSION

In summary, breast cancer research remains a crucial area where technology plays a key role in reducing mortality rates. Despite the development of numerous machine learning (ML) algorithms for analyzing medical datasets, achieving both accuracy and effectiveness in classifying breast cancer data remains a significant challenge. To address this, we proposed a model for breast cancer classification within our web app. Utilizing ML classification techniques such as Decision Tree , K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest , Naïve Bayes , and Logistic Regression , along with ensemble techniques on the WDBC dataset, we conducted a thorough comparison. Although random forest and decision trees exhibit high accuracy on training data, the risk of overfitting diminishes their reliability in real-world scenarios. On the other hand, SVMs with the Gini index criterion strike a balance between complexity and generalization, offering superior accuracy without succumbing to overfitting. Therefore, SVMs appear as a more robust and dependable choice for predictive modeling tasks, particularly when faced with the issues of overfitting.

## 7. FUTURE WORK

Machine learning in medicine enables the application of computational algorithms to analyze medical data, enabling insights into disease diagnosis, prognosis, treatment optimization, and patient consequences [14]. Moving forward, our focus will be on fine tuning our model's performance through hyperparameter tuning, ensuring continuous improvement and better diagnostic capabilities. The evaluation of support vector machine kernel functions for breast cancer prediction assesses the effectiveness of different mathematical transformations in accurately classifying tumor data, providing insights into optimal model performance [15]. The future work will focus on exploring more of the dataset values and yielding more interesting outcomes. This study can help in making more effective and reliable disease prediction and diagnostic system which will contribute towards developing better healthcare system by reducing overall cost, time and mortality rate.

## REFERENCES

[1] Wolberg, William H. "Breast Cancer Wisconsin (Original) Data Set." University of Wisconsin, 1995.

[2] Fielding et al. 1992; Cochran 1997; Burke et al. 2005

[3] Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. Cancer statistics, 2005. CA: a cancer journal for clinicians. 2005 Jan 1;55(1):10-30.

[4] Miller, K., & Jones, P. "Fine Needle Aspiration Cytology in Breast Lesions: A Practical Guide." American Journal of Pathology.

[5] Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications. 2009 Mar 1;36(2):3240-7.

[6] Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007 Jul 1;17(4):694- 701.

[7] Yeh WC, Chang WW, Chung YY. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. Expert Systems with Applications. 2009 May 1;36(4):8204-11.

[8] Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications. 2011 Aug 1;38(8):9573-9.

[9] Kaya Y, Uyar M. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. Applied Soft Computing. 2013 Aug 1;13(8):3429-38.

[10] "Comparison of Different Kernel Functions for SVM in Breast Cancer Prediction" [Journal Article] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1-37.

[11] Libsvm: A Library for Support Vector Machines [Journal Article] Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.

[12] Tang Y. Deep learning using linear support vector machines. *arXiv preprint 1306.0239.* 2013

[13] Vapnik, 1982; Cortes and Vapnik 1995; Duda et al. 2001

[14] Machine Learning in Medicine: A Complete Overview [Journal Article] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Corrado, G. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

[15] Evaluation of Support Vector Machine Kernel Functions for Breast Cancer Prediction [Journal Article] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000).

[16] UCI repository link of data set – https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Created by William Wolberg, Olvi Mangasarian, Nick Street, W. Street

[17] Kaggle licensed data set of UCI –

[18] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data