


# Advances in Breast Cancer Prediction: Evaluating KNN, XGB, and SVM Methods

Qazi Waqas Khan\*

Department of Computer Engineering, Jeju National University, Jeju 63243, Jeju Special Self-Governing Province, Republic of Korea

\*Corresponding author: Qazi Waqas Khan, Department of Computer Engineering, Jeju National University, Jeju 63243, Jeju Special Self-Governing Province, Republic of Korea.

Submitted: 19 July 2024 Accepted: 22 July 2024 Published: 29 July 2024

 <https://doi.org/10.63620/MKSSJCR.2024.1012>

**Citation:** Khan, Q. W. (2024). *Advances in Breast Cancer Prediction: Evaluating KNN, XGB, and SVM Methods*. *Sci Set J of Cancer Res*, 3(3), 01-03.

## Abstract

The number of deaths due to breast cancer is rapidly increasing annually, making it a typical cancer type and primary cause of mortality in females worldwide. Advances in predicting and detecting cancer are crucial for maintaining good health and improving patient care and survival rates. Therefore, achieving a high accuracy rate in cancer prediction is essential for updating the treatment protocols and improving the patient's survival rates. Machine learning approaches have become an important research focus in this area, showing effectiveness in predicting and detecting breast cancer. This study used the K Nearest Neighbor (KNN), Xtreme Gradient Boosting (XGB), and Support Vector Machine (SVM) methods for the classification of cancer patients. The experimental results show that the SVM and XGB models have better prediction results.

**Keywords:** K Nearest Neighbor (KNN), Xtreme Gradient Boosting (XGB), Support Vector Machine (SVM), Cancer Prediction

## Introduction

According to the data from the International Agency for Research on Cancer (IARC), released in December 2020, breast cancer has become the most frequently diagnosed cancer leaving lung cancer behind [1]. Over the past 20 years, from 2000 to 2020, the total number of diagnosed cancer cases has doubled from 10 million to 19.3 million [2]. Nowadays, one in every five people develops cancer at some point in their lives. Research predictions indicate cancer diagnoses will increase by approximately 50% in 2040 [3]. By 2020, the number of deaths due to cancer reached 10 million, compared to 6.2 million in 2000 [4]. As one in every six deaths is due to cancer, highlights the importance of investing in cancer examination and prevention [5]. Information and communication technology (ICT) are integrated into the medical field to successfully transform the healthcare system, specifically in cancer care [6]. The scale and value of data have improved with the help of big data. It has also transformed business intelligence by analyzing diverse, irregular, and incomplete data [7]. It supports prediction and decision-making and is seen as an advancement in raising the standard of patient care while reducing healthcare costs [8]. Due to the effectiveness and accuracy of predicting and detecting diseases, lowering drug prices, and making life-saving decisions in real-time,

data mining algorithms are utilized in the healthcare sector [9]. Our main goal is the successful implementation of different machine-learning classifiers for predicting and diagnosing breast cancer and determining the effectiveness of each classifier by evaluating performance metrics such as confusion matrix, accuracy, precision, and sensitivity.

## Models KNN

KNN, known as k-Nearest Neighbors, is an instance-based learning algorithm for classification and regression [10]. In KNN, data points that are near each other are called neighbors. KNN classifies data points based on the majority class among its "K" nearest neighbors in feature space.

## SVM

SVM is known for its ability to construct the best hyperplane that separates data points from different classes [11]. The data points nearest to the hyperplane are known as support vectors and they determine which hyperplane we should choose. SVM ensures robust classification by maximizing the margin between classes and employs different kernel functions to handle complex non-linear data.

## XGB

XGB is an advanced implementation of gradient boosting designed for supervised learning tasks [12]. It is known for its ef-

fectiveness and high performance in handling large and complex datasets and achieves higher predictive accuracy by sequentially improving weak learners using gradient-boosting methods.

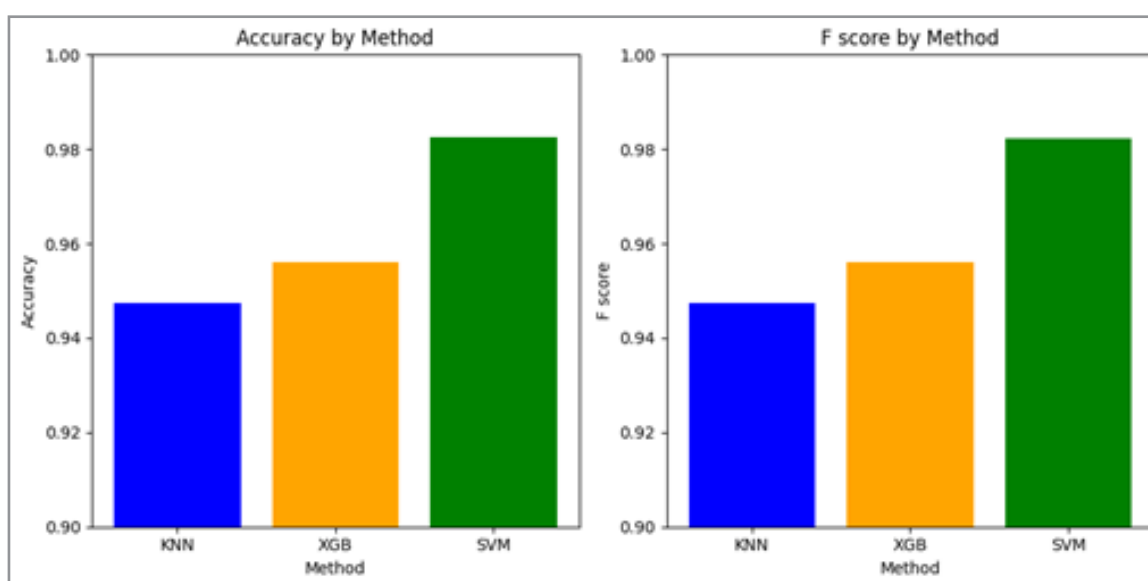
## Experimental Results

**Table 1: Experimental Results of Learning for Cancer Prediction**

Method	Accuracy	Precision	Recall	F score
KNN	0.94742	0.9473	0.9474	0.9473
XGB	0.9561	0.9561	0.9561	0.9560
SVM	0.9825	0.9829	0.9825	0.9824

Table 1 shows the experimental results of KNN, XGB, and SVM for cancer prediction classification without feature selection. Results explain SVM outperforms with high scores across all performance metrics among the classifiers. XGB also performs

well, with an accuracy of 95.61%, while KNN shows the lowest performance among all proposed classifiers. SVM achieves the highest accuracy of 98.25% making itself the most effective and reliable model for cancer prediction.



**Figure 1:** Accuracy and F score of a Learning model for cancer prediction

Figure 1 presents the graphical representation of accuracy and f-score without feature selection across all proposed models. As shown in the figure, SVM achieves the highest prediction accura-

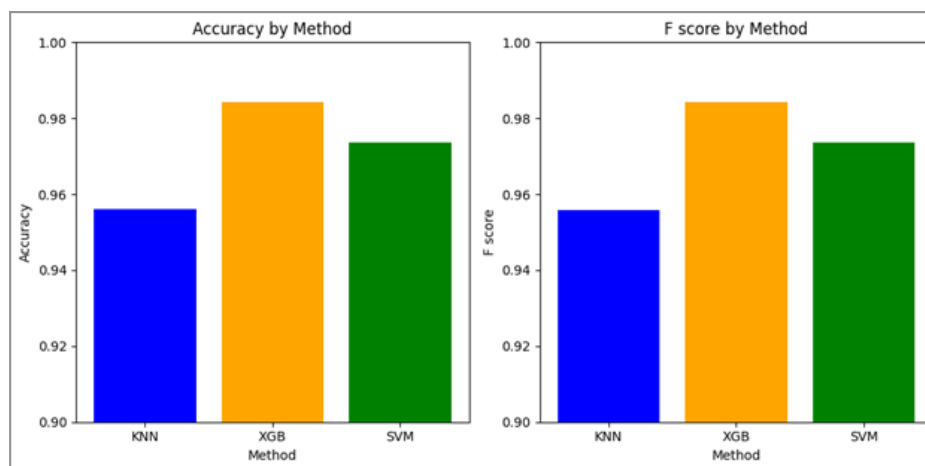
cy and f-score, XGB follows with a slightly lower performance, and KNN shows the least effectiveness among all models.

**Table 2: Experimental Results of Learning for cancer prediction with chi-square feature selection**

Method	Accuracy	Precision	Recall	F score
KNN	0.9561	0.9570	0.9561	0.9559
XGB	0.9842	0.9841	0.9843	0.98419
SVM	0.9737	0.9738	0.9737	0.9736

Table 2 demonstrates the experimental findings of KNN, XGB, and SVM with chi-square feature selection to predict cancer. The results show feature selection method improves the perfor-

mance of KNN and XGB, while the SVM experiences a slight reduction across all evaluation metrics. XGB emerges as the most efficient model with an accuracy of 98.42%.



**Figure 2:** Accuracy and F Score of a Learning Model for Cancer Prediction with Chi-square Feature Selection

The percentage of *Prevotella* spp. in similar studies, that has been performed by the University of Chicago Medical Center is 7,14% in patients with oral cancer and 5,83 in a healthy group. *Porphyromonas* spp (2,07% vs 0,81%, respectively) and *Fusobacterium* (7,67% vs 3,04%, respectively), while *Veillonella* (4,93% vs 7,11%) [10].

Visual representation of accuracy and f-score for the proposed models with chi-square feature selection can be seen in Figure 2. XGB surpasses all other models, achieving the highest accuracy and f-score. SVM also performs well but experiences slightly lower performance scores with the feature selection method, while KNN has the lowest performance.

## Conclusion

Breast cancer prediction is crucial as it is becoming a significant health concern globally due to the rise in cases. This study evaluates three proposed machine learning classifiers and highlights the importance of feature selection in improving the model performance. SVM shows its robustness without feature selection, by achieving the highest accuracy (98.25%), while XGB enhanced its performance with chi-square feature selection, and achieves an accuracy (98.42%). These findings suggest the potential of advanced machine-learning techniques to improve cancer diagnosis and treatment plans.

## References

- Mosteanu, I.-M., Iorga, L.-A., & Mahler, B. (2023). Lung cancer screening—a necessity? –Brief literature review. *Pneumologia*, 71, 188-194.
- Kang, M. J., Jung, K.-W., Bang, S. H., Choi, S. H., Park, E. H., et al. (2023). Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2020. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, 55, 385-399.
- Yan, C., Shan, F., Ying, X., & Li, Z. (2023). Global burden prediction of gastric cancer during demographic transition from 2020 to 2040. *Chinese Medical Journal*, 136, 397-406.
- Ahmad, S. N., Rahimi, A., & Özcebe, H. (2023). Cancer prevalence, incidence, and mortality rates in Afghanistan in 2020: A review study. *Cancer Reports*, 6, e1873.
- Frick, C., Rumgay, H., Vignat, J., Ginsburg, O., Nolte, E., et al. (2023). Quantitative estimates of preventable and treatable deaths from 36 cancers worldwide: a population-based study. *The Lancet Global Health*, 11, 1700-1712.
- Papachristou, N., Kotronoulas, G., Dikaos, N., Allison, S. J., Eleftherochorinou, H., et al. (2023). Digital transformation of cancer care in the era of big data, artificial intelligence and data-driven interventions: navigating the field. *Seminars in Oncology Nursing*, 39(3).
- Turi, J. A., Khwaja, M. G., Tariq, F., & Hameed, A. (2023). The role of big data analytics and organizational agility in improving organizational performance of business processing organizations. *Business Process Management Journal*, 29, 2081-2106.
- Arowoogun, J. O., Babawarun, O., Chidi, R., Adeniyi, A. O., Okolo, C. A., et al. (2024). A comprehensive review of data analytics in healthcare management: Leveraging big data for decision-making. *World Journal of Advanced Research and Reviews*, 21, 1810-1821.
- Ahmed, Z., & Saber, S. (2023). AI-powered analytics in healthcare: enhancing decision-making and efficiency. *International Journal of Applied Health Care Analytics*, 8, 1-16.
- Ukey, N., Yang, Z., Li, B., Zhang, G., Hu, Y., et al. (2023). Survey on exact knn queries over high-dimensional data space. *Sensors*, 23, 629.
- Huajun, W., Li, G., & Wang, Z. (2023). Fast SVM classifier for large-scale classification problems. *Information Sciences*, 642, 119136.
- Kazemi, M. M. K. (2023). Application of XGB-based meta-heuristic techniques for prediction time-to-failure of mining machinery. *Systems and Soft Computing*, 5, 200061