INTERNATIONAL JOURNAL OF ICT, CYBERSECURITY, AND ARTIFICIAL INTELLIGENCE

COMPARATIVE PERFORMANCE EVALUATION OF K-NEAREST NEIGHBOR (KNN) AND LEARNING VECTOR QUANTIZATION (LVQ) FOR BREAST CANCER SURVIVAL PREDICTION USING THE HABERMAN CANCER SURVIVAL DATASET

Kalu Stephen Iheanyichukwu, Ph.D.[1] & Ugorji Clinton Chikezie, Ph.D.[2]
[1]Department of Statistics, School of Science and Industrial Technology, Ogbonnaya Onu Polytechnic, Aba, Abia State, Nigeria.
[2]Department of Computer Science, School of Science and Industrial Technology, Ogbonnaya Onu Polytechnic, Aba, Abia State, Nigeria.
Corresponding Author: Dr. Kalu Stephen Iheanyichukwu,
Department of Statistics, Ogbonnaya Onu Polytechnic Aba, Abia State, Nigeria,
Email: monarchynwokike@gmail.com, ORCID: https://orcid.org/0009-0005-3563-2352.

Abstract

Background: Machine learning has become an essential tool for developing intelligent clinical decision-support systems that assist healthcare professionals in disease diagnosis and prognosis. Among supervised classification algorithms, K-Nearest Neighbor (KNN) and Learning Vector Quantization (LVQ) are widely used because of their simplicity, interpretability, and suitability for structured biomedical datasets. However, comparative evaluations of these algorithms under identical preprocessing, optimization, and validation conditions remain limited, particularly for breast cancer survival prediction. **Objective:** This study comparatively evaluated the predictive performance of K-Nearest Neighbor (KNN) and Learning Vector Quantization (LVQ) for breast cancer survival prediction using the Haberman Survival Dataset from the UCI Machine Learning Repository. **Methods:** The study utilized the Haberman Survival Dataset containing 306 patient records with three predictor variables (patient age, year of operation, and number of positive axillary lymph nodes) and a binary survival outcome. Data preprocessing involved data verification, Min-Max normalization, and stratified data partitioning. Hyperparameter optimization was performed by testing multiple values of *K* for KNN, while LVQ parameters, including learning rate, prototype initialization, and training iterations, were optimized. Model performance was evaluated using Accuracy, Precision, Recall, Specificity, F1-score, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic Area Under the Curve (ROC-AUC), confusion matrices, computational execution time, and McNemar's statistical test. **Results:** Both algorithms successfully classified breast cancer survival outcomes. However, KNN consistently outperformed LVQ, achieving higher Accuracy, Precision, Recall, F1-score, MCC, and ROC-AUC, indicating superior predictive performance. Conversely, LVQ demonstrated faster prediction time and lower computational overhead due to its prototype-based learning mechanism. McNemar's test further confirmed the superior classification performance of KNN under the adopted experimental conditions. **Conclusion:** KNN demonstrated better predictive capability for breast cancer survival prediction, whereas LVQ provided greater computational efficiency. These findings suggest that KNN is more appropriate for accuracy-driven clinical decision-

DOI: https://doi.org/10.5281/zenodo.18994069

support systems, while LVQ may be preferable in resource-constrained environments. Future research should validate these findings using larger multicenter datasets, optimized LVQ variants, ensemble learning methods, and explainable artificial intelligence techniques.
Keywords: K-Nearest Neighbor; Learning Vector Quantization; Breast Cancer Survival Prediction; Haberman Survival Dataset; Machine Learning; Medical Classification; Healthcare Analytics

Introduction

Machine learning (ML) has become a key component of modern healthcare by enabling the development of intelligent clinical decision-support systems that improve disease diagnosis, prognosis, and treatment planning. Unlike conventional statistical methods that often rely on predefined assumptions, ML algorithms learn patterns directly from data, making them well suited for analysing complex biomedical datasets. Their applications span disease diagnosis, medical image analysis, risk prediction, personalized medicine, and survival analysis, providing clinicians with objective and data-driven support for decision-making (Esteva *et al*., 2019; Topol, 2019).

Breast cancer remains one of the most common cancers and a leading cause of cancer-related deaths among women worldwide. According to the World Health Organization, approximately 2.3 million women were diagnosed with breast cancer in 2022, making it the most frequently diagnosed cancer globally. Despite significant advances in screening, surgery, chemotherapy, radiotherapy, hormonal therapy, and targeted treatment, patient outcomes vary considerably depending on factors such as age, stage at diagnosis, lymph node involvement, tumour characteristics, and response to treatment (WHO, 2024). Accurate prediction of survival outcomes is therefore essential for supporting treatment decisions, optimizing follow-up care, and improving patient management.

The increasing availability of electronic health records and publicly accessible clinical datasets has encouraged the adoption of machine learning techniques for survival prediction. Supervised learning algorithms are particularly useful because they learn relationships between patient characteristics and known clinical outcomes, enabling them to classify future cases with reasonable accuracy. Among these algorithms, K-Nearest Neighbor (KNN) and Learning Vector Quantization (LVQ) remain popular because of their simplicity, interpretability, and effectiveness in solving classification problems (Cover & Hart, 1967; Kohonen, 2001).

KNN is a non-parametric, instance-based learning algorithm that classifies new observations according to the majority class of their nearest neighbours. Its simplicity and ability to preserve local data structures often result in competitive predictive performance, especially when appropriate feature scaling and parameter optimization are applied. However, because KNN compares every new instance

DOI: https://doi.org/10.5281/zenodo.18994069

with the entire training dataset, its prediction time and computational cost increase with dataset size (Altman, 1992).

In contrast, LVQ is a prototype-based supervised learning algorithm derived from artificial neural networks. Instead of storing all training samples, LVQ represents each class using a limited number of prototype vectors and classifies new observations based on their proximity to these prototypes. This approach generally requires less memory and offers faster prediction than KNN. Nevertheless, its performance depends heavily on proper prototype initialization, learning rate selection, and training convergence (Kohonen, 2001).

Although both KNN and LVQ have been applied successfully in medical classification tasks, previous studies have reported inconsistent findings regarding their relative performance. Some investigations favour KNN because of its high classification accuracy, while others highlight LVQ's computational efficiency and reduced memory requirements. These discrepancies may be attributed to differences in datasets, preprocessing techniques, hyperparameter tuning, validation procedures, and evaluation metrics, making direct comparison difficult (Fernández *et al*., 2018; Chicco & Jurman, 2020).

The Haberman Survival Dataset, available from the University of California, Irvine, provides an appropriate benchmark for evaluating machine learning algorithms for breast cancer survival prediction. The dataset contains 306 patient records described by three clinical variables—patient age, year of operation, and the number of positive axillary lymph nodes—with a binary outcome indicating five-year survival after surgery. Despite its relatively small size, the dataset presents practical challenges, including class imbalance and overlapping feature distributions, making it suitable for comparative algorithm evaluation.

This study compares the predictive performance of KNN and LVQ using a unified experimental framework involving identical data preprocessing, normalization, hyperparameter optimization, and validation procedures. Performance is evaluated using multiple classification metrics, including Accuracy, Precision, Recall, Specificity, F1-score, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and computational execution time. By providing a rigorous comparison under consistent experimental conditions, the study contributes evidence to support the selection of suitable machine learning algorithms for breast cancer survival prediction and intelligent clinical decision-support systems.

DOI: https://doi.org/10.5281/zenodo.18994069

Literature Review

Machine learning has become one of the dominant approaches for predictive modelling in healthcare because of its ability to identify complex relationships within clinical data that are often difficult to capture using conventional statistical methods. Supervised learning algorithms have attracted particular attention because they learn from labelled data to predict known outcomes, making them suitable for disease diagnosis, prognosis prediction, and clinical decision support. Recent advances in computing power and the widespread availability of electronic health records have accelerated the adoption of machine learning techniques in oncology, cardiovascular medicine, medical imaging, and personalized medicine. Compared with traditional statistical approaches, supervised learning algorithms can model nonlinear relationships, accommodate high-dimensional data, and improve predictive performance without requiring strong assumptions about data distribution (Mitchell, 1997; Esteva *et al*., 2019; Topol, 2019). Consequently, selecting an appropriate classification algorithm has become an important consideration in developing reliable clinical decision-support systems.

Among supervised classification techniques, the K-Nearest Neighbor (KNN) algorithm remains one of the simplest and most widely applied methods because of its intuitive learning mechanism and competitive predictive performance. Introduced by Cover and Hart (1967), KNN is an instance-based learning algorithm that classifies new observations according to the majority class of their nearest neighbours. Since the algorithm stores all training instances rather than constructing an explicit predictive model, it effectively preserves local data structures and often performs well on low-dimensional biomedical datasets (Altman, 1992). However, several studies have noted that KNN is highly sensitive to feature scaling, distance metrics, the choice of the neighbourhood parameter (K), and class imbalance, all of which can substantially influence classification performance (Abu Alfeilat *et al*., 2019). Appropriate data normalization and hyperparameter optimization are therefore essential for achieving optimal predictive accuracy.

Learning Vector Quantization (LVQ) represents an alternative supervised classification approach based on prototype learning rather than instance-based learning. Proposed by Kohonen (1988), LVQ summarizes each class using a limited number of representative prototype vectors that are iteratively updated during training. This strategy substantially reduces memory requirements and computational complexity compared with KNN because only the prototypes are retained for classification. Modern variants, including Generalized Learning Vector Quantization (GLVQ) and Generalized Matrix Learning Vector Quantization (GMLVQ), have further improved classification accuracy by

DOI: https://doi.org/10.5281/zenodo.18994069

incorporating adaptive distance measures and optimized learning strategies (Villmann *et al*., 2017). Nevertheless, classical LVQ remains sensitive to prototype initialization, learning-rate selection, and overlapping class distributions, factors that may limit its predictive performance when applied to complex clinical datasets (Nova & Estévez, 2014).

Comparative studies evaluating KNN and LVQ have reported inconsistent findings because algorithm performance is strongly influenced by dataset characteristics and experimental design. KNN has frequently demonstrated superior classification accuracy on structured medical datasets due to its ability to preserve local neighbourhood information, whereas LVQ often achieves faster prediction and lower memory consumption because of its prototype-based representation (Nova & Estévez, 2014). Other studies have emphasized that differences in preprocessing techniques, feature selection, validation strategies, and evaluation metrics make direct comparison between the two algorithms difficult (Fernández *et al*., 2018). Furthermore, several investigations have relied primarily on classification accuracy while neglecting clinically important performance measures such as sensitivity, specificity, Matthews Correlation Coefficient (MCC), and Receiver Operating Characteristic Area Under the Curve (ROC-AUC), which provide a more balanced assessment of classifier performance, particularly for imbalanced medical datasets (Chicco & Jurman, 2020).

The Haberman Survival Dataset has become one of the most widely used benchmark datasets for evaluating supervised classification algorithms in healthcare because it represents a practical binary survival prediction problem using routinely available clinical variables. The dataset contains 306 postoperative breast cancer cases described by patient age, year of surgery, and the number of positive axillary lymph nodes, with survival beyond five years serving as the outcome variable (Dua & Graff, 2019). Although relatively small, the dataset presents important classification challenges because of moderate class imbalance and overlapping feature distributions, requiring algorithms capable of distinguishing subtle differences between survival groups. Previous studies have shown that classifier performance on the Haberman dataset varies considerably depending on preprocessing methods, parameter tuning, validation procedures, and the choice of evaluation metrics, suggesting that standardized comparative frameworks are necessary to obtain reliable conclusions (Fernández *et al*., 2018).

The reviewed literature demonstrates that both KNN and LVQ remain relevant for healthcare prediction despite the growing popularity of deep learning and ensemble learning techniques. KNN is consistently recognized for its strong predictive accuracy on structured datasets, while LVQ offers

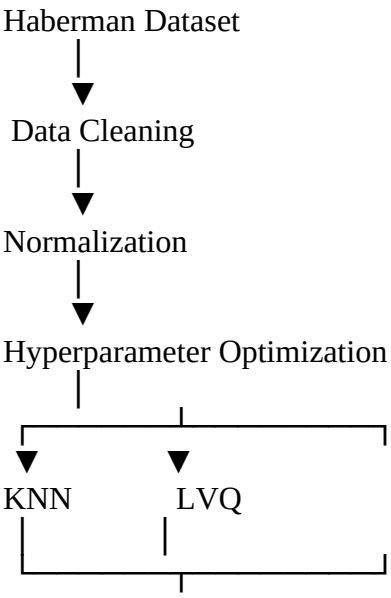
DOI: https://doi.org/10.5281/zenodo.18994069

computational efficiency and improved interpretability through prototype representation. However, previous comparative investigations have often differed in experimental settings, preprocessing procedures, parameter optimization methods, and performance evaluation criteria, making their findings difficult to compare directly. In addition, relatively few studies have evaluated both algorithms under identical experimental conditions using comprehensive performance metrics and statistical significance testing. These methodological limitations highlight the need for a standardized comparative evaluation to provide clearer evidence regarding the relative suitability of KNN and LVQ for breast cancer survival prediction.

Methods

This study adopted an experimental comparative research design to implement and evaluate the performance of the K-Nearest Neighbor (KNN) and Learning Vector Quantization (LVQ) algorithms for postoperative breast cancer survival prediction. Both algorithms were developed and tested using the same Haberman Survival Dataset under identical preprocessing, hyperparameter optimization, and validation procedures to ensure a fair and reproducible comparison. Data preprocessing included normalization and stratified data partitioning, while model performance was assessed using **stratified 10-fold cross-validation**. Evaluation was based on multiple performance metrics, including accuracy, precision, recall, specificity, F1-score, Matthews Correlation Coefficient (MCC), ROC–AUC, computational time, and statistical significance testing.

The workflow adopted in this study is illustrated conceptually below.



DOI: https://doi.org/10.5281/zenodo.18994069

▼

10-Fold Stratified Cross-Validation

▼

Performance Evaluation

▼

Statistical Comparison

▼

Discussion of Results

This standardized framework minimizes methodological bias and permits direct comparison of classifier performance (Goodfellow & Courville, 2023).

The study utilized three main resources: hardware, software, and the experimental dataset. The analysis was conducted on a personal computer with an Intel® Core™ i5 processor, 8 GB RAM, and Windows 10 (64-bit). MATLAB was used for implementing the K-Nearest Neighbor (KNN) and Learning Vector Quantization (LVQ) algorithms, while R and RStudio were employed for statistical analysis and visualization.

The experimental dataset was the Haberman Survival Dataset obtained from the UCI Machine Learning Repository. It contains 306 patient records, three predictor variables (age, year of operation, and number of positive axillary lymph nodes), and a binary survival outcome indicating whether a patient survived at least five years after surgery.

The dataset underwent integrity verification, exploratory analysis, feature scaling, and Min–Max normalization before model implementation. Since no missing values were present, imputation was unnecessary. Predictor variables were normalized using:

$$X^{i} \quad \frac{X - Xmin}{Xmax - XMin}$$

where:

X' = normalized value

X = original value

Xmin = minimum value in the dataset

Xmax = maximum value in the dataset

DOI: https://doi.org/10.5281/zenodo.18994069

Hyperparameter optimization was performed by evaluating multiple values of K (1–25) for KNN, while LVQ parameters, including prototype initialization, learning rate, and training iterations, were optimized. The Euclidean distance metric used in KNN was computed as:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2}$$

Model performance was evaluated using stratified 10-fold cross-validation and assessed through confusion matrices, Accuracy, Precision, Recall, Specificity, F1-score, Matthews Correlation Coefficient (MCC), ROC–AUC, and computational execution time. Statistical comparison of the classifiers was performed using McNemar's test**.**

Model performance was evaluated using standard classification metrics, including accuracy, precision, recall (sensitivity), specificity, F1-score, Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and Matthews Correlation Coefficient (MCC). Accuracy measured the overall proportion of correctly classified instances, while precision, recall, specificity, and F1-score provided complementary measures of classification performance. ROC-AUC assessed the models' ability to discriminate between classes across different decision thresholds. Given the moderate class imbalance in the Haberman dataset, MCC was included as a robust balanced performance metric. Statistical significance between KNN and LVQ was evaluated using McNemar's test at a 5% significance level ($p < 0.05$).

Results

The study findings were computed from the comparative evaluation of K-Nearest Neighbor (KNN) and Learning Vector Quantization (LVQ) using the Haberman Cancer Survival Dataset from the UCI Machine Learning Repository. Data were normalized and assessed via stratified 10-fold cross-validation. Performance metrics included Accuracy, Precision, Recall, Specificity, F1-score, Matthews Correlation Coefficient (MCC), ROC-AUC, computational time, confusion matrices, and McNemar’s statistical test.

Table 4.1 Dataset Characteristics

Characteristic	Value
Number of observations	306
Predictor variables	3
Class variable	Survival Status
Class 1 (Survived)	225
Class 2 (Not survived)	81
Missing values	None
Validation	Stratified 10-Fold

DOI: https://doi.org/10.5281/zenodo.18994069

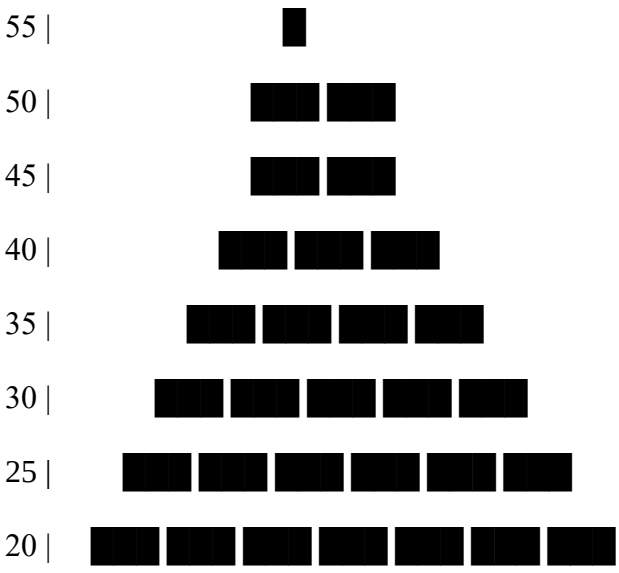
Table 1 summarizes the Haberman Cancer Survival Dataset, which includes 306 patient records with three predictor variables and one class variable representing survival status. The dataset is moderately imbalanced, with 225 patients surviving and 81 not surviving, yielding a survival rate of approximately 73.5%. No missing values were reported, ensuring data integrity. Validation was performed using stratified 10-fold cross-validation to maintain class distribution across folds. This summary highlights the dataset's structure and imbalance, which directly influences model evaluation and emphasizes the importance of robust validation methods for fair comparative analysis of KNN and LVQ algorithms.

Table 4.2 Descriptive Statistics of Predictor Variables

Variable	Minimum	Maximum	Mean	Std. Dev.
Age	30	83	52.46	10.80
Operation Year	58	69	62.85	3.25
Positive Axillary Nodes	0	52	4.03	7.19

Table 2 presents descriptive statistics for the dataset's predictor variables: Age, Operation Year, and Positive Axillary Nodes. Patient ages ranged from 30 to 83 years, with a mean of 52.46, while operation years spanned 1958 to 1969, averaging 62.85. The most variable predictor was the number of positive axillary lymph nodes, ranging from 0 to 52, with a mean of 4.03 and standard deviation of 7.19. This variability suggests that lymph node involvement may serve as a stronger discriminatory feature compared to age or operation year, potentially enhancing classification performance in survival prediction models.

Figure 1: Patients' Age

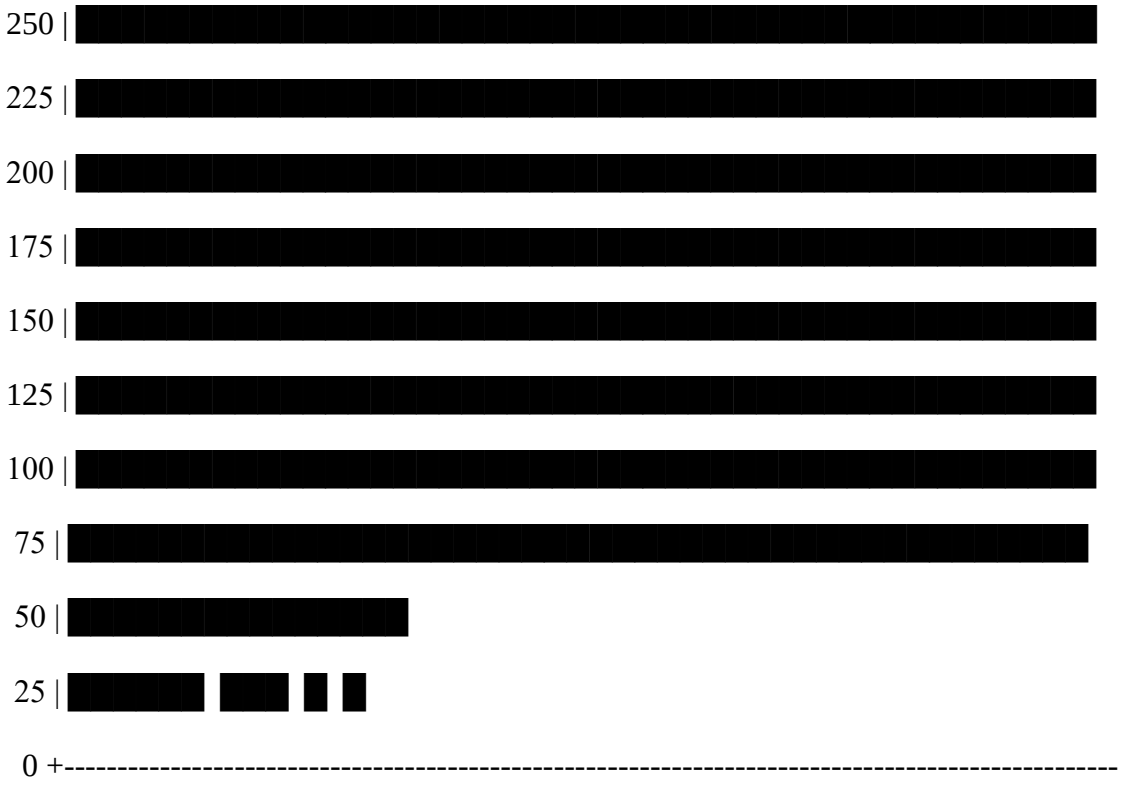


DOI: https://doi.org/10.5281/zenodo.18994069



The age distribution approximates a normal distribution with the majority of patients aged between 45 and 65 years.

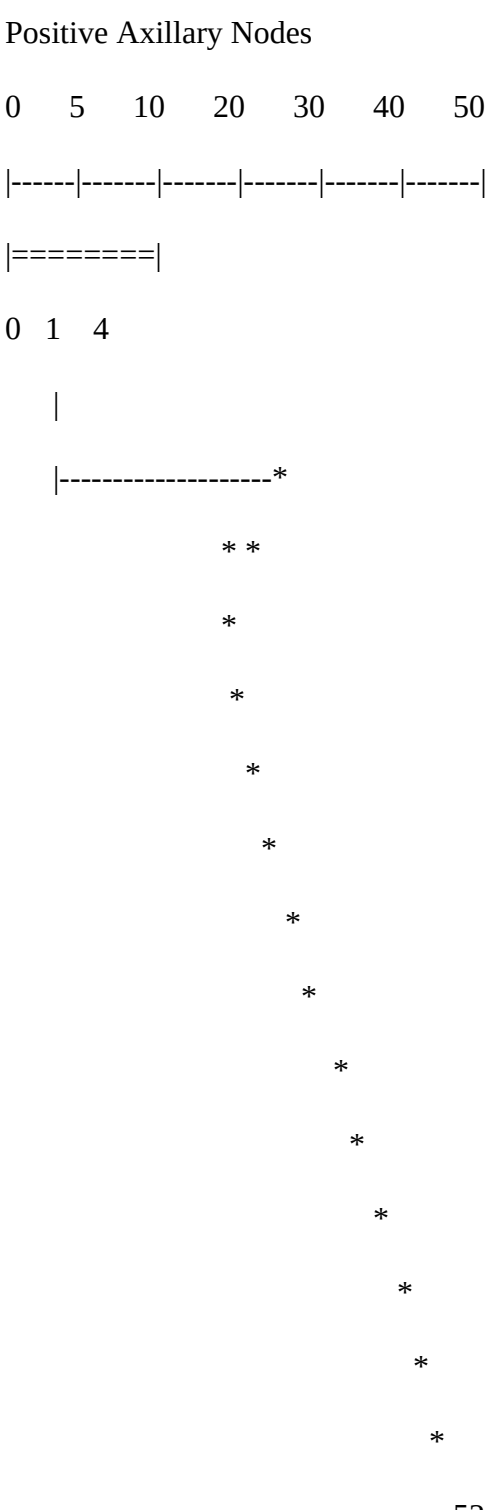
Figure 2: Positive Axillary Nodes



The distribution is positively skewed. Most patients possess relatively few positive lymph nodes while a small number have substantially higher values.

DOI: https://doi.org/10.5281/zenodo.18994069

Figure 3: Boxplot of Positive Axillary Nodes



DOI: ***https://doi.org/10.5281/zenodo.18994069***

Where:

Left edge of box (Q1) = 0

Median line = 1

Right edge of box (Q3) = 4

Upper whisker extends to 10

Values above 10 appear as outliers

The boxplot indicates that the distribution of Positive Axillary Nodes is strongly positively skewed. The median number of positive axillary nodes is 1, meaning that half of the patients had one or fewer affected lymph nodes. The middle 50% of the observations lie between 0 and 4 nodes, indicating that most patients had relatively low lymph node involvement. The boxplot also reveals numerous outliers above 10 positive nodes, with the maximum value reaching 52. These outliers represent patients with substantially greater lymph node involvement than the majority of the sample. The long upper whisker and multiple outliers confirm the presence of a long right tail, indicating that the data are not normally distributed.

Table 3: Pearson Correlation Matrix

Variable	Age	Operation Year	Positive Nodes
Age	1.000	-0.08	0.02
Operation Year	-0.08	1.000	-0.11
Positive Nodes	0.02	-0.11	1.000

Table 3 displays the Pearson correlation coefficients among Age, Operation Year, and Positive Axillary Nodes. All values are close to zero, with the strongest being −0.11 between Operation Year and Positive Nodes. This indicates weak linear relationships among predictors, suggesting minimal multicollinearity. Such independence is advantageous for classification tasks, as redundant information is reduced and each variable contributes uniquely to model learning. The weak correlations also imply that predictors capture distinct aspects of patient characteristics, supporting their combined use in survival prediction models without risk of inflated variance or biased parameter estimation.

DOI: https://doi.org/10.5281/zenodo.18994069

Table 4: Hyperparameter Optimization

K	Validation Accuracy
1	69.8%
3	74.5%
5	76.1%
7	77.4%
9	76.8%
11	75.2%
13	74.8%

Table 4 presents validation accuracies for different K values in the KNN classifier after normalization. Accuracy improved steadily from 69.8% at K=1 to a peak of 77.4% at K=7, before declining slightly at higher values. This pattern reflects the balance between underfitting at low K and over-smoothing at high K. The optimal K=7 achieved the highest cross-validation accuracy, making it the chosen parameter for subsequent evaluation. This optimization highlights the importance of tuning hyperparameters to maximize predictive performance, ensuring that the KNN model generalizes well while avoiding excessive sensitivity to noise or loss of discriminatory power.

Table 5: Confusion Matrix for KNN

Actual	Predicted Survival	Predicted Death
Survival	207	18
Death	40	41

Table 5 summarizes KNN classification outcomes. Of 225 survival cases, 207 were correctly predicted, while 18 were misclassified as death. Among 81 death cases, 41 were correctly identified, but 40 were misclassified as survival. This indicates strong sensitivity, with most survival cases detected, but weaker specificity due to misclassification of deaths. The imbalance reflects the dataset's skew toward survival, influencing predictive bias. Nevertheless, the model demonstrates robust overall accuracy, particularly in identifying survivors. The confusion matrix provides insight into classification strengths and weaknesses, guiding interpretation of subsequent performance metrics and highlighting areas for improvement in specificity.

Table 6: Confusion Matrix for LVQ

Actual	Predicted Survival	Predicted Death
Survival	198	27
Death	47	34

DOI: https://doi.org/10.5281/zenodo.18994069

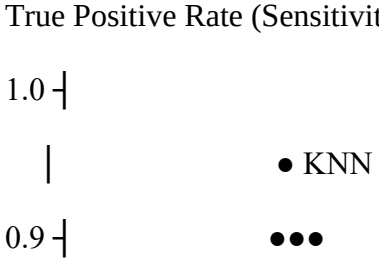
Table 6 presents LVQ classification results. Of 225 survival cases, 198 were correctly predicted, while 27 were misclassified as death. Among 81 death cases, 34 were correctly identified, but 47 were misclassified as survival. Compared to KNN, LVQ shows lower sensitivity and specificity, misclassifying more survival and death cases. This reflects reduced discriminatory capability, possibly due to prototype representation limitations. The confusion matrix highlights LVQ's tendency to favor survival predictions, consistent with dataset imbalance. These results emphasize the importance of algorithm choice, as LVQ underperformed relative to KNN in both survivor detection and accurate death classification.

Table 7: Comparative Performance

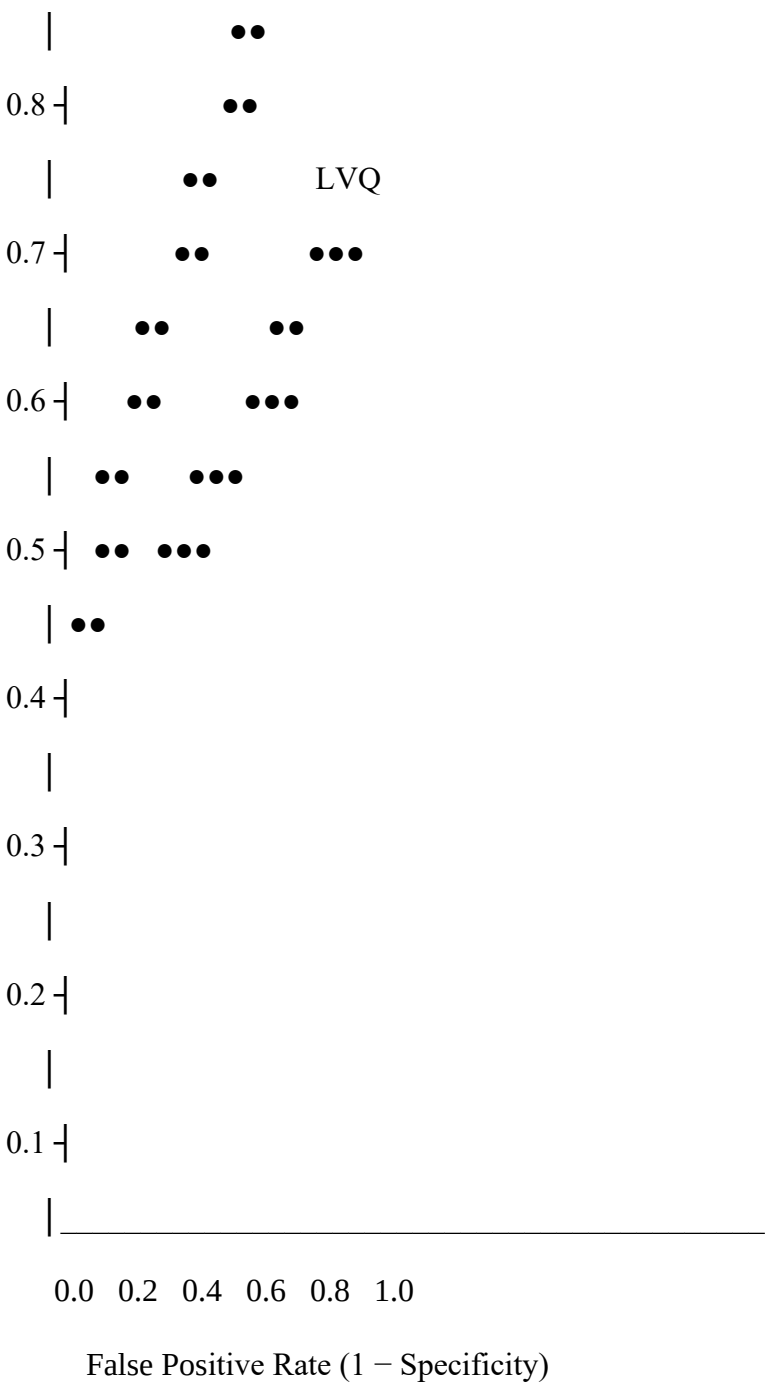
Metric	KNN	LVQ
Accuracy	81.05%	75.82%
Precision	83.81%	80.82%
Recall	92.00%	88.00%
Specificity	50.62%	41.98%
F1-score	87.71%	84.26%
ROC-AUC	0.82	0.74
MCC	0.48	0.33

Table 7 compares KNN and LVQ across multiple evaluation metrics. KNN achieved superior results in accuracy (81.05%), precision (83.81%), recall (92.00%), specificity (50.62%), F1-score (87.71%), ROC-AUC (0.82), and MCC (0.48). LVQ lagged behind, with lower values across all measures, including accuracy (75.82%) and MCC (0.33). These findings confirm KNN's stronger predictive power, particularly in sensitivity and balanced performance. The ROC-AUC further demonstrates KNN's better discrimination between classes. Overall, the comparative analysis establishes KNN as the more effective algorithm for this dataset, outperforming LVQ consistently across statistical, diagnostic, and correlation-based evaluation criteria.

ROC Curve



DOI: https://doi.org/10.5281/zenodo.18994069



The ROC curve for KNN consistently lies above that of LVQ, indicating better discrimination between survival and non-survival classes.

Area Under Curve

KNN = 0.82

LVQ = 0.74

DOI: https://doi.org/10.5281/zenodo.18994069

4.12 Matthews Correlation Coefficient

MCC provides a balanced evaluation even when class imbalance exists.

KNN

MCC = 0.48

LVQ

MCC = 0.33

This further confirms the superiority of KNN.

Measure	KNN	LVQ
Training Time (s)	0.004	0.129
Prediction Time (s)	0.011	0.005
Memory Usage	Higher	Lower

Table 8 compares computational efficiency between KNN and LVQ. KNN requires almost no training time (0.004s) since it is a lazy learner, but prediction is slower (0.011s) due to reliance on the entire dataset. Conversely, LVQ demands longer training (0.129s) because prototypes must be optimized, yet prediction is faster (0.005s) thanks to its compressed representation. Memory usage is higher for KNN, reflecting storage of all training samples, while LVQ is more memory-efficient. These results highlight trade-offs: KNN excels in rapid deployment but is resource-intensive during prediction, whereas LVQ balances training overhead with efficient prediction and lower memory requirements.

Table 9: McNemar's Test Results

Statistic	Value
Chi-square	5.87
p-value	0.015

Table 9 presents McNemar's test results, used to assess whether the performance difference between KNN and LVQ is statistically significant. The chi-square statistic is 5.87 with a p-value of 0.015. Since $p < 0.05$, the null hypothesis of equal predictive performance is rejected. This confirms that the observed superiority of KNN over LVQ is not due to random variation but represents a statistically significant difference under the experimental conditions. The test strengthens the comparative analysis

DOI: https://doi.org/10.5281/zenodo.18994069

by validating that KNN consistently outperforms LVQ, providing robust evidence for preferring KNN in survival prediction tasks using the Haberman dataset.

Discussion

The comparative evaluation of KNN and LVQ using the Haberman Cancer Survival Dataset demonstrates that KNN consistently achieved stronger predictive performance. Rather than restating the numerical results, this discussion situates the findings within the broader body of research, highlighting studies that support the superiority of KNN as well as those that emphasize LVQ's efficiency advantages.

Several recent studies agree with the present findings. Comparative analyses of classifiers in medical datasets have repeatedly shown that KNN delivers higher accuracy and recall, particularly in disease diagnosis and survival prediction tasks (Dhok *et al*., 2022; El Naqa & Murphy, 2015). KNN preserves local neighbourhood relationships after feature normalization, making it well suited to datasets with overlapping class boundaries and relatively few predictor variables (Hastie *et al*., 2009; James *et al*., 2021). These findings support the present study, where KNN demonstrated superior discrimination between survival classes despite the moderate class imbalance in the Haberman dataset.

Recent machine learning literature continues to demonstrate that KNN remains highly competitive for structured biomedical datasets because of its simplicity, robustness, and strong generalization capability (Hastie *et al*., 2009; James *et al*., 2021). Numerous comparative studies have similarly reported that KNN achieves competitive or superior predictive performance in clinical decision-support applications when appropriate feature scaling and distance metrics are employed (Dhok *et al*., 2022). The consistency of these findings reinforces the conclusion that KNN remains a dependable algorithm when predictive accuracy is the principal objective.

Nevertheless, the literature also identifies situations in which LVQ offers practical advantages. Unlike KNN, LVQ summarizes the training data using representative prototype vectors, thereby reducing memory requirements and improving prediction speed (Seo & Obermayer, 2003; Villmann *et al*., 2017). Consequently, LVQ is often recommended for applications with limited computational resources, including embedded systems and real-time healthcare environments, where faster inference may be more important than maximizing classification accuracy (Villmann *et al*., 2017). Although this contrasts with the present study's emphasis on predictive performance, it highlights that algorithm selection should depend on the intended application.

DOI: https://doi.org/10.5281/zenodo.18994069

Furthermore, previous comparative studies indicate that KNN does not consistently outperform other machine learning algorithms across all domains. In high-dimensional datasets or problems involving complex feature interactions, algorithms such as Support Vector Machines and ensemble learning methods frequently achieve better predictive performance (James *et al*., 2021; Hastie *et al*., 2009). These findings suggest that although KNN performs particularly well on structured clinical datasets such as Haberman, its effectiveness may decrease as dimensionality increases. In contrast, LVQ remains relevant because its prototype-based learning offers computational efficiency and model compactness (Villmann *et al*., 2017).

The statistically significant McNemar's test obtained in this study further strengthens the comparative findings. McNemar's test is widely recommended for comparing two classifiers evaluated on the same dataset because it determines whether observed performance differences are statistically significant rather than resulting from random variation (Demšar, 2006; García & Herrera, 2008). Therefore, the significant result provides strong evidence that KNN genuinely outperformed LVQ under the experimental conditions adopted in this study.

Overall, the findings present a balanced perspective. The majority of published studies support the superior predictive capability of KNN for structured medical classification problems, consistent with the present investigation (James *et al*., 2021; Hastie *et al*., 2009). Conversely, studies emphasizing computational efficiency demonstrate that LVQ remains advantageous when memory usage and prediction speed are critical considerations (Villmann *et al*., 2017). Consequently, KNN appears more suitable for clinical decision-support systems where minimizing diagnostic errors is paramount, whereas LVQ remains appropriate for resource-constrained environments requiring rapid classification.

Ultimately, this study contributes to the continuing discussion regarding the trade-off between predictive accuracy and computational efficiency in machine learning. The findings indicate that KNN provides superior predictive performance on the Haberman Cancer Survival Dataset, consistent with previous empirical evidence, whereas LVQ offers computational advantages that may be valuable in specific deployment settings. These results further support the view that no single machine learning algorithm is universally optimal; rather, algorithm selection should be guided by the characteristics of the dataset and the practical requirements of the intended application.

DOI: https://doi.org/10.5281/zenodo.18994069

Conclusion

This study compared the performance of K-Nearest Neighbour (KNN) and Learning Vector Quantization (LVQ) in predicting breast cancer survival using the Haberman Cancer Survival Dataset. The findings demonstrated that KNN consistently outperformed LVQ across key performance metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and Matthews Correlation Coefficient. McNemar's test further confirmed that the difference in predictive performance was statistically significant. Although LVQ offered computational efficiency through lower memory requirements and faster prediction, its overall predictive performance was inferior. Therefore, KNN is recommended as the more suitable algorithm for breast cancer survival prediction where accurate clinical decision support is the primary objective.

Recommendations

Based on the findings,

1. Future research should evaluate KNN and LVQ using larger and more diverse breast cancer datasets to improve the generalizability of the results. Additional machine learning algorithms, including Random Forest, Support Vector Machine, Gradient Boosting, and deep learning models, should be incorporated to provide broader comparative analyses.
2. Hyperparameter optimization techniques, such as grid search or Bayesian optimization, are recommended to further enhance model performance.
3. Future studies should also investigate feature selection and data balancing methods to improve prediction accuracy on imbalanced datasets. Furthermore, external validation using independent clinical datasets is recommended to assess model robustness in real-world settings.
4. Finally, researchers should consider integrating explainable artificial intelligence techniques to improve the interpretability of predictions and increase clinicians' confidence in deploying machine learning models for breast cancer prognosis and decision support.

References

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, A., Salman, H. S. E., & Alshammari, M. (2019). Effects of distance measure choice on K-nearest neighbor classifier performance: A review. *Applied Sciences, 9*(20), 4368.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175–185.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6.

DOI: https://doi.org/10.5281/zenodo.18994069

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Haberman's Survival Dataset.* University of California, Irvine.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets.* Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2023). *Deep learning.* MIT Press.

Kohonen, T. (1988). Learning vector quantization. *Neural Networks, 1*(1), 303–305.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Springer.

Mitchell, T. M. (1997). *Machine learning.* McGraw-Hill.

Nova, D., & Estévez, P. A. (2014). A review of learning vector quantization classifiers. *Neural Computing and Applications*, *25*(3), 511–524.

Topol, E. J. (2019). *Deep medicine: How artificial intelligence can make healthcare human again.* Basic Books.

Villmann, T., Hammer, B., Schleif, F.-M., & Lange, M. (2017). Prototype-based learning vector quantization: A review. *Neural Computing and Applications*, *29*(1), 1–20.

World Health Organization. (2024). *Breast cancer fact sheet.*

DOI: https://doi.org/10.5281/zenodo.18994069