

Innovative K-nearest neighbour-based breast cancer classification using robust feature selection and data balancing techniques

Saad Ullah^a, Muhammad Nadeem^{b,*}, Zohaib Mushtaq^c, Shafqat Ali^d

^a Department of Electrical, Electronics & Computer Systems, University of Sargodha.

^b Department of Electrical, Electronics & Computer Systems, University of Sargodha.

^c Department of Electrical, Electronics & Computer Systems, University of Sargodha.

^d University Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi.

* Corresponding author: Author Muhammad Nadeem, Email: nadeem.abdullah@uos.edu.pk

Received: 10 February 2025, Revised: 24 September 2025, Accepted: 29 September 2025, Published: 01 October 2025

KEY WORDS

KNN
SMOTE
CHI2
L1 regularization
Stratified K-fold cross-validation

ABSTRACT

Cancer is one of the most formidable health conditions worldwide, and it is often difficult to diagnose in its early stages. More than 10 million people die in a year due to this disease. Breast cancer is one of the most lethal disorders of women, among others. The research presented here is devoted to the use of machine learning algorithms to identify breast cancer at an early stage with minimal errors. In particular, the K-Nearest Neighbour (KNN) algorithm was employed, along with several distance functions, class balancing implemented using SMOTE (Synthetic Minority Over-sampling Technique), and a stratified K-fold cross-validation approach, ensuring that the data were distributed evenly. Also, CHI2 and L1 feature selection algorithms had been employed to improve the results. The accuracy rate of the study is 97.96% which is better than the related work done recently.

1. Introduction

Cancer is now a more global issue, and the diagnoses of breast cancer have risen over the past 10 years. Early detection is key to the best possible treatment and survival [1], [2]. Breast cancer is the most common cancer in women worldwide. Breast cancer is diagnosed in about 2300 thousand people each year, which is more than the number of lung cancers being discovered [3]. Before this, breast cancer was diagnosed by several different ways: physical examination, mammography, and fine needle aspiration or core needle biopsy[4]. Breast cancer was often detected through mammography, with an 87% to 99% specificity [5].

Machine learning (ML) has also become a ubiquitous part of cancer diagnostics, such as breast cancer, and is

transforming the field of medicine. The number of women in the world with breast cancer is up from 10 years ago, but the disease continues to rise. Early detection offers the best chance of successful treatment and long-term survival [6]. A similar study and comparison of multiple Machine Learning algorithms, such as logistic regression, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbours, Support Vector Machine, and AdaBoost, was described previously [7].

Machine learning has proved effective throughout the process of cancer recognition by healthcare providers due to the usage of wrapper-based feature selection and three methods: LR, LSVM, and QSVM. These are the machine learning algorithms that we have

compared (with and without feature selection) with the WBC (Wisconsin Breast Cancer) and WDBC (Wisconsin Diagnostic Breast Cancer) datasets to predict cancer. It has been determined that the LR classifier using feature selection is significantly better on accuracy of, 97.1 and 83.5 with WDBC and WPBC, respectively [8]. The K-Nearest Neighbours (KNN) algorithm is one of the most popular breast cancer detection algorithms. It has seen significant application to medical diagnostics when testing on datasets such as the Wisconsin Breast Cancer and Wisconsin Diagnostic Breast Cancer datasets of the UC Irvine Machine Learning Repository because of its extremely easy-to-use and highly effective nature[8]. In this research, we aim to evaluate the accuracy of diagnosing breast cancer using a machine learning model that incorporates the pipelining concept. The pipeline uses SMOTE, stratified k-fold cross-validation, and K-Nearest Neighbours. This study responds to these issues in the following main ways:

Pipelining is used because simple k-fold cross-validation will yield incorrect results if it selects folds that exclude malignant variants. To counter this, we implement stratified k-fold cross-validation, which provides equal class distribution and representation across the entire dataset, thus mitigating the risk of overfitting. To account for this vast difference between benign and malignant patients, SMOTE is used to create a balanced dataset.

The KNN algorithm is used to compute performance metrics, such as statistical testing and confusion matrix statistics, to distinguish between cancerous and benign cases. The model in question yields exceptional and precise results. While most research has focused on just one or two distance functions, this study employs multiple distance functions and proposes a technique that is applicable across different distance metrics.

In the paper, Section 2 presents the related work. The proposed methodology is provided in Section 3. The fourth section shares the experimental results, and then the fifth section closes the paper.

2. Related Work

Previous studies show that classification is an important factor for diagnosis in diseases such as thyroid disease, heart disease, lung cancer, and breast cancer [9]. The literature shows that breast cancer, compared with other illnesses, demands a smarter diagnosis[5]. The basic aim of a literature review is to identify a problem with existing approaches and propose solutions.

In article [10], Mushtaq et al. Applied the Naïve Bayes Machine Learning Method and the feature selection methods K-PCA, Sigmoid, and Chi2 to identify breast cancer with high sensitivity. As mentioned in the article [4], the paper suggested six machine learning algorithms, including Gaussian Naïve Bayes, decision trees, and support vector machines. They produced more accurate five-year breast cancer risk estimates than BCRAT tools alone, which may lead to increased early detection and prevention.

In Table 1, a literature overview is presented, including methodologies and accuracy outcomes from various studies. In line with the work described above, we are comparing evaluation measures in the Wisconsin Breast Cancer dataset for diagnosis. Data analysis must achieve the highest accuracy with the lowest error rate. Table 1 summarizes the methods in the literature.

Table 1

Summary of existing studies reported in the literature

Ref	Approaches	Feature Selection	Data set	Evaluation Metrics	Accuracy
[3]	KNN, SMOTE	-	WBC	AUC	95.32%
[2]	SVM, KNN	LSVC, Chi2 Square	WBC	AUC	97.14%
[4]	SVM	K-means algorithm	WBC	AUC	98.06%
[5]	Naïve Bayes, J48,	10-fold cross-validation	WBC	AUC	97.27%
			WBC	AUC(J48)	94.26%
			WBC	AUC	97.4%
[6]	RBL-RBFNN	-	WBC		97.0%
			D		

3. Methodology

Fig. 1 - The hierarchy and sequence of the proposed approach to breast cancer diagnosis. This approach uses the Wisconsin Breast Cancer dataset in the UCI (University of California, Irvine) repository. The first crucial step is preprocessing, which removes useless or redundant features and handles missing values by using the mean of the entries of the involved feature.

Several benign and malignant patients fluctuate through time. To correct for this disparity, we implement the SMOTE oversampling procedure using the pipelining principle. The best parameters are calculated using stratified K-fold cross-validation on the training data, ensuring an equal distribution and relation from the entire dataset. On the training and test data, the K-Nearest Neighbour algorithm was used for prediction and modelling.

- $U_{pn} = U_p + U_{negg} = \text{total occurrence containing } U$
- $U_{nn} = U_p^- + U_{negg}^- = \text{total occurrence not including } U$

Hence,

$$U_p^- = \frac{(U_p + U_p^-) + (U_p + U_{negg})}{U_n} \quad (2)$$

$$U^2 = \sum_{m=1}^n \frac{(O_m - E_m)^2}{E_m} \quad (3)$$

where U^2 denotes the chi-square statistic, O is the observed value, and E represents the expected value.

$$U^2 = \frac{U_n((U_p \times U_{negg}) - (U_{negg} \times U_p^-))^2}{(U_p + U_p^-)(U_{negg} + U_{negg}^-)(U_{pn})(U_{nn})} \quad (4)$$

Figure 2 highlights the features selected after preprocessing and scaling the dataset. Initially, the WBC dataset had 11 columns; however, after removing the ID number and class columns, it was reduced to 9 columns. Additionally, two further features were discarded by applying Chi-square feature selection, as they showed no association with the class variables. Using L1-based feature selection, we removed three features that were not as important for predicting the target variable.

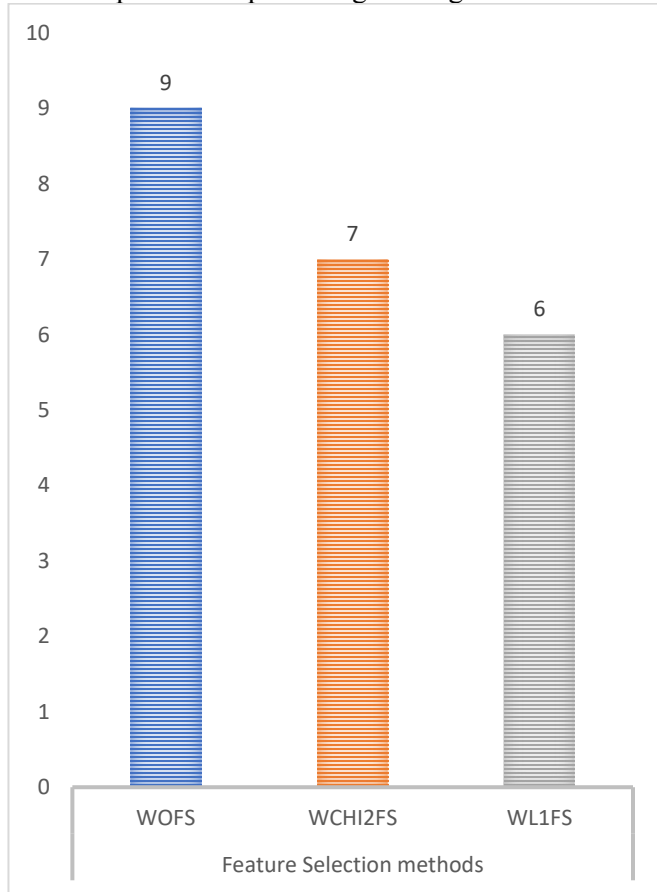


Fig. 2. Number of features selected in techniques

3.5 Pipelining Strategy

The application of SMOTE (Synthetic Minority Over-sampling Technique), stratified K-fold cross-

validation, and the KNN algorithm into one pipeline may provide a powerful predictive model of a classification problem, especially in cases when dealing with an imbalanced dataset.

3.5.1 SMOTE

The development of SMOTE was needed to overcome the problem of insufficient minority class examples in a dataset. It addresses the issue of unbalanced data, where there is unequal representation of target class categories, by creating other samples as indicated in Figure 3. SMOTE is applied to each minority class by using k-nearest neighbours to increase the amount of data in that category to a sufficient strength to balance the majority.

The dataset comprises 458 Benign cases and 241 Malignant cases. Since benign cases are approximately twice as numerous as malignant cases, class imbalance becomes a concern that may affect accuracy in Figure 3. To address this imbalance, SMOTE is utilized in this study to balance the class distribution, thereby enhancing the predictive model's performance.

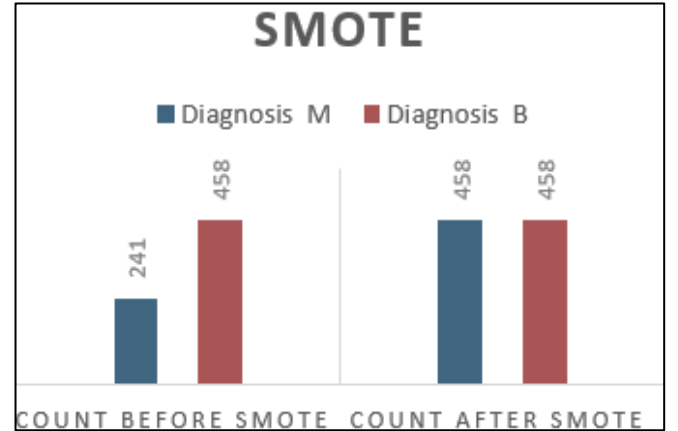


Fig. 3 Balance the class distribution using SMOTE

3.5.2 Stratified K-fold Cross Validation

In this study, the stratified k-fold cross-validation method is employed instead of simple k-fold cross-validation. The reason for this choice is that simple k-fold cross-validation might produce inaccurate results, especially if some folds happen to lack malignant cases entirely. This could lead to misleading performance metrics, as the model might not be adequately evaluated on all possible classes.

3.5.3 K-Nearest Neighbours (KNN) algorithm

The K-Nearest Neighbours (KNN) algorithm is a straightforward, adaptable, and widely utilized method for classification. Various distance functions are applied in this approach, with common ones including Euclidean distance, Manhattan distance, and Minkowski distance. The selection of a distance metric can significantly

impact the algorithm's effectiveness[2]. Suppose 'u' and 'w' be the feature vector, where $u = u_1, u_2, u_3 \dots, u_n$ and $w = w_1, w_2, w_3 \dots, w_n$ —the mathematical expression of various distance functions and under.

$$\text{Minkowski } d(u,w) = \left(\sum_{j=1}^n |u_j - w_j|^k \right)^{\frac{1}{k}} \quad (5)$$

where $k=1,2,3 \dots \infty$

$$\text{Euclidean } d(u,w) = \sqrt{\sum_{j=1}^n (u_j - w_j)^2} \quad (6)$$

$$\text{Manhattan } d(u,w) = \sum_{j=1}^n |u_j - w_j| \quad (7)$$

$$\text{Hamming } d(u,w) = \begin{cases} 0 & \text{if } u = w \\ 1 & \text{if } u \neq w \end{cases} \quad (8)$$

$$\text{Canberra } d(u,w) = \sum_{j=1}^n \frac{|u_j - w_j|}{|u_j| + |w_j|} \quad (9)$$

$$\text{chebyshev } d(u,w) = \max_{j=1}^n |u_j - w_j| \quad (10)$$

3.6 Performance Evaluation Metrics

In KNN, performance evaluation metrics are used to determine how well the model is functioning. The confusion matrix compares actual and predicted classes, with True Positive (α) representing the number of positive instances correctly identified. However, misclassification occurs with β , where a True Negative is incorrectly labelled, leading to a False Positive and indicating a misdiagnosis. Similarly, True Negative is incorrectly categorized as False Negative. To evaluate the efficiency of the proposed technique, the following formulas are applied.

True Positive Rate

$$TRP = \frac{\alpha}{\alpha + \beta} \quad (11)$$

True Negative Rate

$$TNR = \frac{\beta}{\beta + \alpha} \quad (12)$$

False Positive Rate

$$FPR = \frac{\alpha}{\beta + \alpha} \quad (13)$$

False Negative Rate (FNR)

$$FNP = \frac{\beta}{\alpha + \beta} \quad (14)$$

Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (15)$$

Precision

$$\text{Precision} = \frac{TP}{TP+FP} \quad (16)$$

Recall

$$\text{Recall} = \frac{TP}{TP+FN} \quad (17)$$

F1 Score

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (18)$$

Matthews Correlation Coefficient (MCC)

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (19)$$

Error Rate

$$\text{Error Rate} = \frac{FP+FN}{TP+TN+FP+FN} \quad (20)$$

4. Experimental Results and Discussion

The performance of the KNN algorithm was assessed in the current study using the stratified K-fold cross-validation. The rationale of applying stratified K-fold is to ensure that the distribution of the data following the up-sampling is maintained as opposed to the simple K-fold cross-validation. Eight kinds of distance functions have been treated in the KNN algorithm; the K value being ranged within 1 to 60. The process is done thrice. On the given WBC data, we would first not make use of any feature selection techniques, which is denoted by WOFS. However, in the second iteration, we used the CHI2 method for feature selection. In CHI2, we are utilizing the seven most prominent features for cancer detection. Lastly, we have used L1-based regularization for feature selection, through which we have selected six features to train and test our algorithm. The efficiency of different functions is shown in Figures 4, 5, and 6.

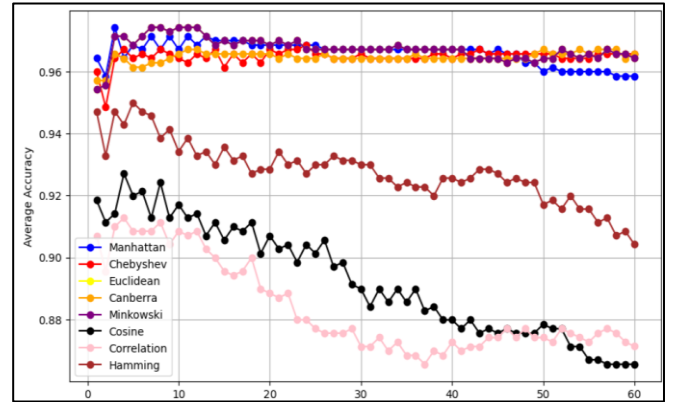


Fig.4 KNN Accuracy for Different k Values and Distance Metrics using k-fold cross-validation WOFS

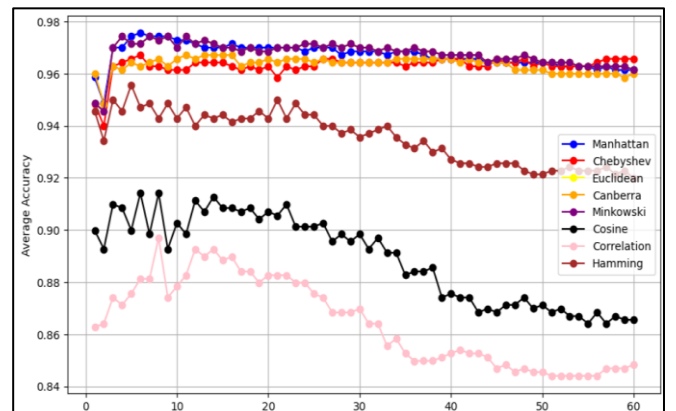


Fig.5 KNN Accuracy for Different k Values and Distance Metrics using k-fold cross-validation with CHI2 feature selection

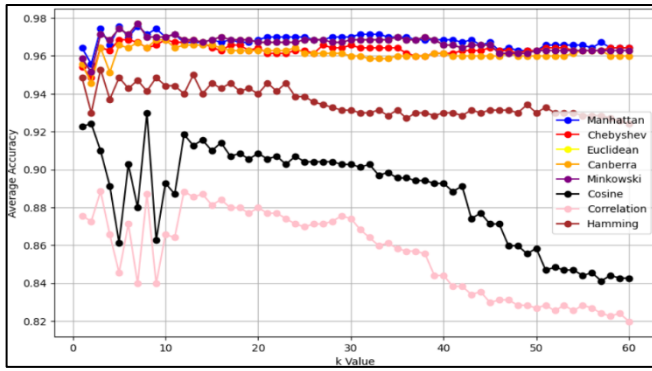


Fig. 6 KNN Accuracy for Different k Values and Distance Metrics using k-fold cross-validation with L1 regularization feature selection

The ROC curves for the above distance functions based on WOFS, CHI2, and L1 regularization are given in Figures 7, 8, and 9. It can be observed from Figure 4 that, in the case without feature selection, the highest efficiency is achieved by Manhattan, Euclidean, and Minkowski, which is 97.42 for K=3, K=7, K=7 respectively. The least efficiency of 88.84 at K=3 is achieved by correlation. Furthermore, for CHI2 feature selection, we have efficiencies of 97.56, 97.42, and 97.42 for Manhattan, Euclidean, and Minkowski, with values of K = 6, K = 7, and K = 7, respectively. The least efficient value of 89.69 at K = 8 is achieved through correlation. Lastly, the efficiencies obtained by L1 regularization feature selection are 97.96, 97.71, and 97.71 for Manhattan, Euclidean, and Minkowski, respectively, at K=7, K=7, and K=7. The least efficiency of 88.84% and 92.98% at K=3 and K=8 is achieved by correlation and Cosine.

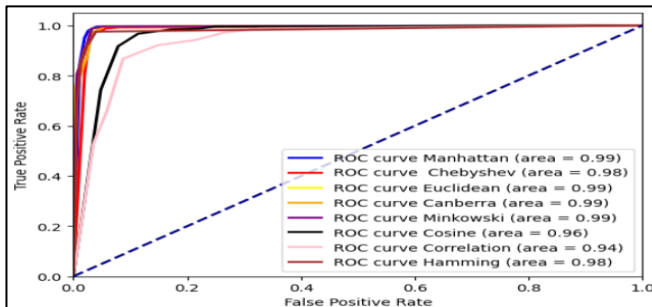


Fig.7. Receiver Operating Characteristic (ROC) WOFS

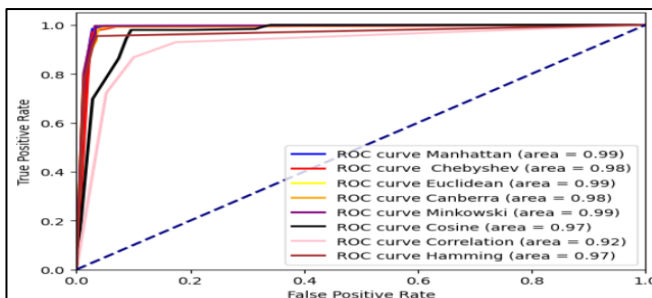


Fig. 8 Receiver Operating Characteristic with CHI2 feature selection

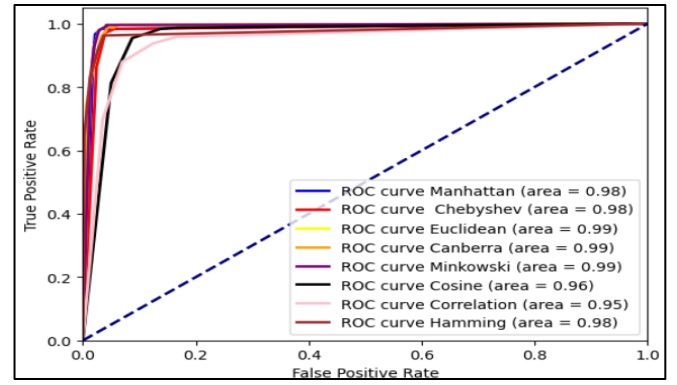


Fig. 9. Receiver Operating Characteristic with L1 regularization feature selection

The other parameter used to evaluate the binary classifier is AUC (Area Under the Curve). Its value varies from 0 to 1. The performance indexes are generated using a confusion matrix for the values of K in the KNN algorithm, where maximum efficiency is achieved.

Table 3

Performance Evaluation Metrics of WOFS

		Without Feature Selection (WOFS)						
Performance Evaluation Metrics		Accuracy %	Recall %	Fail-out%	F1-score%	Error-Rate%	MCC%	Miss-rate%
KNN-distance functions with best K values range from 1 to 60.	Manhattan K=3	97.42	96.68	2.18	96.28	2.57	94.31	3.31
	Chebyshev K=24	96.42	97.09	3.93	94.92	3.57	92.22	2.90
	Euclidean K=7	97.42	97.92	2.83	96.32	2.57	94.37	2.07
	Canberra K=12	96.70	98.34	4.14	95.37	3.29	92.92	1.65
	Minkowski K=7	97.42	97.92	2.83	96.32	2.57	94.37	2.07
	Cosine K=4	92.70	95.43	8.73	90.01	7.29	84.63	4.56
	Correlation K=4	91.27	87.96	6.98	87.42	8.72	80.74	12.03
	Hamming K=5	94.99	90.04	2.40	92.53	5.00	88.85	9.95

Table 3 shows that the Euclidean method at K = 7 and the Minkowski method at K = 7 are tied as the best classifiers, offering balanced performance across all metrics, including a low miss rate and the highest MCC. If our priority is to prefer the MCC rate, then the above distance functions are best. However, if the aim is to reduce false positives, we should divert our effort towards the Manhattan distance at K=3.

Table 4

Performance Evaluation Metrics of CHI2 Feature Selection

		Feature Selection Based on CHI2 (WFS)						
Performance Evaluation Metrics		Accuracy %	Recall %	Fall-out%	F1-score %	Error-Rate%	MCC%	Miss-rate%
KNN-distance functions with best K values range from 1 to 60	Manhattan K=6	97.56	97.92	2.62	96.52	2.43	94.67	2.07
	Chebyshev K=6	96.70	97.51	3.71	95.33	3.29	92.85	2.48
	Euclidean K=7	97.42	98.75	3.27	96.35	2.57	94.43	1.24
	Canberra K=11	96.70	98.75	4.36	95.39	3.29	92.96	1.24
	Minkowski K=7	97.42	98.75	3.27	96.35	2.57	94.43	1.24
	Cosine K=6	91.41	96.68	11.35	88.59	8.58	82.52	3.31
	Correlation K=8	89.69	86.72	8.73	85.30	10.30	77.40	13.27
	Hamming K=5	95.56	90.87	1.96	93.39	4.43	90.13	9.12

Table 4 indicates that Manhattan at K=6 and Euclidean at K=7 have the best performance across multiple metrics. Manhattan has the attribute of having the lowest fallout, which ensures a minimum false positive rate. Euclidean, on the other hand, has a minimum miss-rate, making it the most suitable candidate where false negatives are crucial.

Table 5

Performance Evaluation Metrics of L1-Based Regularization Feature Selection

		Feature Selection Based on L1 Regularization (WFS)						
Performance Evaluation Metrics		Accuracy %	Recall %	Fall-out%	F1-score %	Error-Rate%	MCC%	Miss-rate%
KNN-distance functions with best K values range from 1 to 60.	Manhattan K=7	97.96	98.34	2.83	96.53	2.43	94.70	1.65
	Chebyshev K=6	96.70	97.92	3.93	95.35	3.29	92.88	2.07
	Euclidean K=7	97.71	99.58	3.27	96.77	2.28	95.09	0.41
	Canberra K=9	96.85	98.34	3.93	95.56	3.14	93.21	1.65
	Minkowski K=7	97.71	99.58	3.27	96.77	2.28	95.09	0.41
	Cosine K=8	92.98	97.92	9.60	90.59	7.01	85.66	2.07
	Correlation K=3	88.84	86.72	10.04	84.27	11.15	75.07	13.27
	Hamming K=3	95.27	91.70	2.83	93.05	4.72	89.50	8.29

Table 5 shows that the Manhattan method at K = 7 and the Euclidean method at K = 7 deliver the best performance across multiple metrics. However, Manhattan has a slight advantage in terms of better performance in the F1 score region. Table 6 compares our results with those of the current leading approaches and techniques. Our method achieved the highest

accuracy when compared to other studies on the identical dataset.

Table 6

Comparison of the proposed accuracy with other relevant studies on the same dataset

Year [Ref]	Accuracy	Approaches	Data set
2020 [2]	97.14 %	SVM, KNN	WBC
2022 [3]	95.32 %	KNN, SMOTE	WBC
2023[8]	97.10 %	LR, LSVM	WBC
This Study	97.96 %	SMOTE, Stratified K-fold CV, KNN	WBC

5. Conclusion

In this study, we aimed to determine the optimal distance function for the KNN algorithm while addressing the challenges of imbalanced datasets. By combining SMOTE for data balancing and stratified K-fold cross-validation to maintain proper distribution, we ensured that our classifier was trained on a well-represented dataset rather than one that artificially inflates accuracy. Without proper distribution, standard K-fold cross-validation can sometimes yield misleadingly high accuracy because some data folds may miss important instances, such as true negatives. While K-fold works well on its own, when SMOTE is applied, stratified K-fold cross-validation is necessary to maintain data integrity.

It was observed that the Euclidean distance function with seven neighbouring entities (K = 7), combined with L1 regularization, achieved the highest accuracy of 97.71%. In comparison, the Manhattan distance function with six neighbouring entities (K = 6), along with CHI² feature selection, achieved an accuracy of 97.42%. Furthermore, the Manhattan distance function with K = 3 also achieved an accuracy of 97.42%.

These findings highlight that selecting the proper distance function and feature selection method can significantly impact performance, ensuring a well-balanced and reliable classifier rather than one that appears accurate but is not.

6. References

- [1] R. Kazerani, "Improving Breast Cancer Diagnosis Accuracy by Particle Swarm Optimization Feature Selection," *International Journal of Computational Intelligence Systems*,

- vol. 17, no. 1, pp. 1–30, Dec. 2024, doi: 10.1007/S44196-024-00428-5/TABLES/18.
- [2] Z. Mushtaq, A. Yaqub, S. Sani, and A. Khalid, “Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets,” *Journal of the Chinese Institute of Engineers, Transactions of the Chinese Institute of Engineers, Series A*, vol. 43, no. 1, pp. 80–92, Jan. 2020, doi: 10.1080/02533839.2019.1676658.
- [3] S. Motarwar and D. K. Jha, “Assessment of Factors Influencing the Survival of Breast Cancer Patients using a Machine Learning Approach,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 11, no. 3, pp. 80–84, Jan. 2022, doi: 10.35940/ijitee.C9713.0111322.
- [4] W. L. Al-Yaseen, A. Jehad, Q. A. Abed, and A. K. Idrees, “The Use of Modified K-Means Algorithm to Enhance the Performance of Support Vector Machine in Classifying Breast Cancer,” *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 2, p. 190, 2021, doi: 10.22266/ijies2021.0430.17.
- [5] M. T. Ahmed, M. N. Imtiaz, and A. Karmakar, “Analysis of Wisconsin Breast Cancer original dataset using data mining and machine learning algorithms for breast cancer prediction,” *Journal of Science Technology and Environment Informatics*, vol. 9, no. 2, pp. 665–672, 2020, doi: 10.18801/jstei.090220.67.
- [6] A. H. Osman and H. M. A. Aljahdali, “An Effective of Ensemble Boosting Learning Method for Breast Cancer Virtual Screening Using Neural Network Model,” *IEEE Access*, vol. 8, pp. 39165–39174, 2020, doi: 10.1109/ACCESS.2020.2976149.
- [7] M. Yuvalı, B. Yaman, and Ö. Tosun, “Classification Comparison of Machine Learning Algorithms Using Two Independent CAD Datasets,” *Mathematics 2022, Vol. 10, Page 311*, vol. 10, no. 3, p. 311, Jan. 2022, doi: 10.3390/MATH10030311.
- [8] R. Hasan and A. S. M. Shafi, “Feature Selection based Breast Cancer Prediction,” *International Journal of Image, Graphics and Signal Processing*, vol. 15, no. 2, pp. 13–23, Apr. 2023, doi: 10.5815/IJIGSP.2023.02.02.
- [9] S. Khalil *et al.*, “Enhancing Ductal Carcinoma Classification Using Transfer Learning with 3D U-Net Models in Breast Cancer Imaging,” *Applied Sciences (Switzerland)*, vol. 13, no. 7, Apr. 2023, doi: 10.3390/app13074255.
- [10] Z. Mushtaq, M. F. Qureshi, M. J. Abbass, and S. M. Q. Al-Fakih, “Effective kernel-principal component analysis based approach for wisconsin breast cancer diagnosis,” *Electron Lett*, vol. 59, no. 2, Jan. 2023, doi: 10.1049/el12.12706.
- [11] “Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository.” Accessed: Aug. 11, 2025. [Online]. Available: <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>