Improving K-Nearest Neighbor Algorithm Performance Using Modified Distance Measures

Ashok Kumar[1] and Deepanshu Mishra[1*]

[1]Department of Statistics, University of Lucknow, Lucknow, Uttar Pradesh, India
***Corresponding author-**
Corresponding author email: deepanshu.mishra2011@gmail.com

Abstract: Classification in the field of machine learning refers to the process of identifying and categorizing objects within a given dataset. Distance-based algorithms are widely used for data classification problems. The k-nearest neighbor (KNN) classification is based on measuring the distances between the test sample and the training samples to determine the final classification output. KNN relies on measuring similarity to group data into classes based on how similar their features are without relying on probabilities but rather utilizing distance metrics, for classification purposes. However, it's important to note that the effectiveness of the KNN algorithm heavily depends on selecting a distance measure. This paper delves into exploring the use of a supervised learning technique known as the k-nearest neighbor (KNN) algorithm for data classification. This paper proposes modified Sorensen and Canberra distance measures designed to enhance the performance of the KNN algorithm by more effectively capturing relationships between data points. The proposed distance measures performance has been evaluated using various datasets, such as the Iris, Breast Cancer, and Diabetes datasets. The performance of the KNN algorithm using the modified distance measures against the original Sorensen, Canberra and most popular measure Euclidean distance has been compared. Results demonstrate that the KNN algorithm employing the proposed modified distance measure consistently outperforms its counterparts in terms of accuracy, precision, recall, and F-score across various datasets. Notably, the modified distance measure exhibits robust performance, even outshining existing distance measures in scenarios involving outlier sensitivity and increased dimensionality.

The paper concludes with insights into the applicability of the modified distance measures beyond KNN, suggesting its potential for enhancing accuracy in various classification tasks.

Keywords: Machine learning, k-nearest neighbor, distance measure, accuracy measures.

1. Introduction

In machine learning, classification refers to the process in which a model tries to categorise objects by identifying and analysing a given dataset. In supervised machine learning, for classifying data the KNN algorithm is one of the popular methods. The KNN is one of the oldest, simplest and most accurate algorithms for pattern classification and regression models.

KNN was proposed in 1951 by [11], and then modified by [9]. Classification algorithms use input training data to identify patterns within the training data and predict the likelihood that the given data that follows will fall into one of the predetermined categories [3]. It is used for classification purposes which uses a similarity measure to classify data in classes based on their feature similarity, without relying on probabilities. This approach assumes that instances with similar features are likely to be located near each other in the dataset [13].

The KNN algorithm has been studied over the past few decades and is widely applied in many fields [18]. Thus, the KNN comprises the baseline classifier in many pattern classification problems such as pattern recognition [16], text categorization [16], ranking models [25], object recognition [4], and event recognition [27] applications.

There are three main components of the KNN algorithm:

(a) The collection of stored objects, which is also known as the training dataset.
(b) A distance or similarity metric that helps determine the similarities between different objects and
(c) An appropriate value for k, which dictates how many nearest neighbors should be considered in making predictions [1].

The KNN algorithm is used for classifying datasets based on distance measures for a training model that closely matches the testing query. The KNN classification algorithm identifies a nearby cluster of k objects in the training dataset that are similar to the target object and then predict the class of these closest objects as the class for the target object [1]. The KNN stands out among various classification algorithms due to its simplicity, widespread usage in classification tasks, and its adaptive and easily comprehensible design [15]. However, its performance heavily depends on the distance measure used to determine the similarity measures between data points. The KNN algorithm used various existing distance measures such as Euclidean, Sorensen and Canberra distance may not always identify the within relationship, for every dataset leading to suboptimal classification outcomes. The primary focus of the KNN classifier or model has been on data sets with pure numerical features [23]. However, the KNN model can also be applied to other types of data including categorical data [7].

This paper aims to propose the modified distance measures for Sorensen and Canberra distances to improve the performance of the KNN algorithm. These modified distance measures are designed to identify the relation between points more effectively, and also, lead

to improve the accuracy of the classification. To evaluate the effectiveness of the proposed distance measures, we compare the performance of the KNN algorithm using the modified distance measures with existing distance measures on Iris, Breast cancer and Diabetes datasets. The result shows that the KNN algorithm with the modified measures outperformed as compared to existing distance measures including Euclidian distance measure. The findings suggest that the proposed modified distance measures improve the performance of the KNN algorithm. The proposed approach offers an avenue for improving the accuracy of various classification problems.

Section 2 briefly describes the existing distance measure and proposed distance measures. Error and performance measures are also given in this section. Section 3 explains the dataset used and the experimental analysis of the datasets using the KNN model with modified and original distance measures are presented. The comparison of the performance of the KNN model on the proposed modified measure with different existing distance measures is also discussed in this section. The conclusion of the study is presented in Section 4.

2. Distance and Error Measures

In this section, various distance and similarity measures have been described. The performance measures for evaluating the performance of the model are also given.

2.1 Distance and Similarity Measures

Distance measure plays a vital role in many machine-learning tasks such as classification, clustering, and anomaly detection. Choosing a distance measure is an important aspect of clustering the data because it determines how the similarity between two objects is observed. This decision has an impact on the structure of the clusters, as certain elements may be considered close to each other based on one distance metric but far apart according to another [18]. By considering the most similar samples among their nearest neighbors, the k-nearest neighbors algorithm identifies the class of an unlabelled test sample.

A specific distance measure is used for calculating the distance between each training data sample and the test sample [2].

The distance measure is a function that determines the similarity and dissimilarity between two data points. It is a numerical value that indicates how close or far two data points are in a given feature space [20]. It is typically a positive real number where a lesser value shows a similarity of a higher degree. The distance function between two vectors x and y is a function $d(x; y)$ that defines the distance between both vectors as a non-negative real number. This function is

considered as a metric if satisfy a certain number of properties [10]. There are many distance measures in machine learning, some of the most common distance measures for continuous data used in this paper are as follows:

(a) Euclidean Distance

The Euclidean distance or Euclidean metric is the ordinary distance between two points that one would measure with a ruler. This distance represents the root of the sum of the squares of differences between the opposite values in vectors. In n dimensions, the Euclidean distance between two points $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined by:

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

where x_i and y_i are the coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$ in the i^{th} dimension.

(b) Sorensen Distance

The Sorensen distance [21], also known as Bray-Curtis is one of the most commonly applied measurements to express relationships between the feature values of $\boldsymbol{x}$ and $\boldsymbol{y}$. For n dimensions, it is computed as the ratio of the sum of the absolute difference and the sum of the corresponding features of two data points and is defined as:

$$D(x,y) = \frac{\sum_{i=1}^{n}|x_i - y_i|}{\sum_{i=1}^{n}(x_i + y_i)}$$

where x_i and y_i are the coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$ in the i^{th} dimension.

(c) Canberra Distance

Canberra distance is introduced by [24] and modified in [14]. It is the sum of the absolute difference between the corresponding features of two data points divided by the sum of the absolute value of the corresponding feature values prior to summing and is defined as

$$D(x,y) = \sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|}$$

where x_i and y_i are the coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$ in the i^{th} dimension.

Existing distance measures may not be suitable for all types of data or problems. There are several drawbacks of existing distance measures including its sensitivity to outlier data points, issues of performance decrement as a result of a large number of features and assumptions about data. For example, the Euclidean distance, which is one of the most common distance measures, can be sensitive to outliers and may not be able to capture the similarity between data points that are not linearly separable. Also, as the number of features increases in a dataset, the performance of distance measures is impacted. This problem is also called the *curse of dimensionality* [8].

The Sorensen distance is helpful for comparing the makeup of different groups, but it has some important drawbacks. It can be overly affected by zeros and large values, making it less reliable when the groups have significant differences in less common elements. It doesn't work with negative values and can give misleading results when the overall size of the groups matters. The measure also struggles with high-dimensional data, where it becomes less effective, and it doesn't fit well with techniques that expect distances to behave like traditional Euclidean distances [6].

The Canberra distance measure has several drawbacks, primarily due to its extreme sensitivity to small values and zeros. It can produce disproportionately large distances when one or both of the vector elements are close to zero, which can skew results in datasets with many small or zero values. This sensitivity to small changes makes it unstable and less reliable for noisy data. Additionally, it is not well-suited for high-dimensional data, as the distance measure becomes less meaningful when there are many features. The computational complexity also increases with larger datasets, making it less efficient compared to simpler distance measures like Euclidean [6].

In such a scenario, there is a need to propose new or modified distance measures that can provide better accuracy than existing ones for the classification and clustering of the features. Modified distance measures would be helpful in improving the performance of the classification models.

In this paper, we are proposing the modified distance measures for Sorensen and Canberra distance measures. The proposed distance measures are as given:

- **Modified Sorensen Distance Measures:** The expression for modified Sorensen distance measure is given by

$$d(x,y) = \left(\frac{\sum_{i=1}^{n}|x_i - y_i|^3}{\sum_{i=1}^{n}(x_i + y_i)^3}\right)^{1/3}$$

  where x_i and y_i are the coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$ in the i^{th} dimension.

- **Modified Canberra Distance Measure:** The expression for modified Sorensen distance measure is given by

$$d(x,y) = \left(\sum_{i=1}^{n}\frac{|x_i - y_i|^3}{|x_i|^3 + |y_i|^3}\right)^{1/3}$$

  where x_i and y_i are the coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$ in the i^{th} dimension.

2.2 Error Measures and Performance Matrix for Classification

Performance evaluation metrics for classification models can produce multiple categorical outputs. Most error measures typically calculate the overall error in our model, but they do not provide visibility into individual instances of errors. For a model, it is possible to misclassify some categories more frequently than others; however, this information cannot be obtained using standard accuracy measures. In addition, in cases where the data has a noticeable class imbalance, meaning that one class has significantly more instances than the other classes, a model may tend to predict the majority class for all cases and result in a high accuracy score but poor performance while predicting the minority classes. Confusion matrices become valuable tools in such cases.

The confusion matrix, also known as the error matrix, is a widely used visualisation method in the field of machine learning for presenting the outcomes and results of models used for classification problems [3,19]. It provides a contingency table layout that allows for an intuitive and concise measurement of the level of confusion within the classification model. Each cell in this matrix represents either correct or incorrect predictions made by the model based on its judgment. By examining each row representing real categories and each column representing predicted labels, one can gain valuable insights into how well the classification model performs in terms of accuracy [3,22]. The components of the confusion matrix are defined in [17] as follows:

- *True Positive (TP):* It is defined as the number of times model the actual positive values are equal to the predicted positive. The model predicts a positive value, which is correct.
- *False Positive (FP)*: It is also known as type I error and is defined as the number of times the model wrongly predicts negative values as positives. The model predicts a negative value, which is actually positive.
- *True Negative (TN):* It is defined as the number of times actual negative values are equal to predicted negative values. The modal predicts a negative value, which is actually negative.
- *False Negative (FN):* It is also known as type II error and is defined as the number of times the model wrongly predicts negative values as positives. The modal predicts a negative value, and it is actually positive.

Some of the following criteria have been used to evaluate the performance of the model defined in [2] as follows:

(a) Accuracy

It tells how often the model classifies a right class. It is defined as the ratio of the number of objects classified correctly to the total number of objects classified.

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

(b) Precision

It is the ability of the model to classify positive values correctly. It is defined as the actual correct prediction divided by the total prediction made by the model.

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

(c) Recall

It is used to calculate the model's ability to predict positive values. It is calculated by the number of true positives divided by the total number of true positives and false negatives.

$$Recall = \frac{TP}{TP + FN}$$

(d) F-Score

It is the harmonic mean of Recall and Precision. It is useful when we need to take both Precision and Recall into account.

$$F - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

(e) AUC Score

Area under the curve (AUC) is the popular metric to evaluate the performance of classifiers. It can take values 0 to 1. A higher AUC score indicates superior performance in the model's capacity to distinguish between positive and negative outcomes.

The performance measures of the proposed modified distance measures will be used to evaluate the performance of the KNN model for the classification of the object.

3. Experimental Analysis and Discussion

This section evaluates the effectiveness of the KNN model with proposed modified and existing similarity measurements over three datasets from different domains. The dataset used for performing the model training in this work was acquired from the UCI ML repository (https://archive.ics.uci.edu/). This repository is one of the most reliable and used dataset sources for researching and implementing machine-learning algorithms. The datasets viz., Iris, Breast cancer and Diabetes are used to evaluate the performance measures of the KNN model. *Iris* is one of the most widely used datasets in machine learning and is available in open source. The dataset contains 50 samples of three species of Iris namely *Setosa*, *Virginica* and

Versicolor. The *Breast cancer* dataset has 569 instances and 30 variables and is also available in open source. The *Diabetes* dataset is also publicly available and has 768 instances and 9 variables. All datasets vary in their characteristics, including the number of features, attributes, and sizes.

All datasets are compared with the modified distance measures with the original distance measures including the Euclidean distance measure. The datasets were analysed using the Python programming language with various packages. For the KNN model algorithm, the value of k was chosen to find the optimal parameter value for a given model using ten-fold cross-validation. The KNN model were trained on two sizes 65% and 80% for all datasets. Performance evaluation of the KNN model based on modified distance measures with original distance measures along with Euclidian distance measure was done using Accuracy, precision, recall, F score and AUC.

(a) The KNN Model performance for Iris Dataset

The KNN model has been trained for the Iris data with 65% training size and 80% training size of the data. The performance measures for 65% of the training size for the iris dataset by applying the KNN model with the difference distance measures are listed in Table 1.

Table 1: Performance Measures of the KNN model for the Iris dataset for 65% training size.

Distance Measure	Accuracy	Precision	Recall	F-Score
Modified Sorensen	0.981	0.980	0.970	0.980
Modified Canberra	0.981	0.980	0.970	0.980
Sorensen	0.981	0.980	0.970	0.980
Canberra	0.962	0.960	0.960	0.960
Euclidian	0.981	0.980	0.970	0.980

From Table 1, it is observed that the performance of the KNN model using the modified Canberra distance measure outperformed the original Canberra distance measure in all performance measures and it is equally good as other distance measures. Although, the KNN model for the modified Sorensen distance measure performed equally well as compared to the original Sorensen and Euclidean distance measure it outperformed than original Canberra distance measures in terms of accuracy, precision, recall and F score. The AUC score for both modified distance measures was 1.

The performance measures for 80% of the training size for the iris dataset for the KNN model with the difference distance measures are listed in Table 2.

Table 2: Performance Measures of the KNN model for the Iris dataset for 80% training size.

Distance Measure	Accuracy	Precision	Recall	F-Score
Modified Sorensen	1.000	1.000	1.000	1.000

Modified Canberra	1.000	1.000	1.000	1.000
Sorensen	0.967	0.980	0.940	0.960
Canberra	0.967	0.980	0.940	0.960
Euclidian	0.967	0.980	0.940	0.960

From Table 2, it is observed that for 80% of the training size for the iris dataset, the performance of the KNN model using both the modified distance measures outperformed its original distance measures along with the Euclidean distance measure. Both the modified distance measures outperformed in terms of accuracy, precision, recall and F score. The AUC score for both modified distance measures was 1. It is also noted that as the training size of the dataset increases the performance measures of the KNN model increase for both the modified distance measures, which shows that the proposed modified distance measures can be used for other classification models too.

(b) The KNN Model performance for the Breast Cancer Dataset

The KNN model has been trained for the Breast cancer dataset with 65% training size and 80% training size of the dataset. The performance measures for 65% of the training size for the KNN model with the difference distance measures are listed in Table 3.

Table 3: Performance Measures of the KNN model for Breast Cancer dataset for 65% training size.

Distance Measure	Accuracy	Precision	Recall	F-Score
Modified Sorensen	0.950	0.950	0.950	0.950
Modified Canberra	0.950	0.950	0.950	0.950
Sorensen	0.950	0.950	0.950	0.950
Canberra	0.955	0.960	0.940	0.950
Euclidian	0.950	0.950	0.950	0.950

For Table 3, it is observed that 65% of the training dataset for the KNN model with Modified distance measures performed equally well as other distance measures. The modified distance measures performed the same in terms of accuracy, precision, recall and F score. The AUC score for both modified distance measures was 0.98.

The performance measures for 80% of the training size for the breast cancer dataset for the KNN model with the difference distance measures are listed in Table 4.

Table 4: Performance Measures of the KNN model for Breast Cancer dataset for 80% training size.

Distance Measure	Accuracy	Precision	Recall	F-Score
Modified Sorensen	1.000	0.940	0.950	0.950
Modified Canberra	1.000	0.940	0.950	0.950
Sorensen	0.967	0.950	0.960	0.960
Canberra	0.967	0.950	0.940	0.950
Euclidian	0.967	0.940	0.950	0.950

It can be seen from Table 4 that for 80% of the training dataset for breast cancer, the KNN model with modified Sorensen distance measure outperformed its original distance measures along with Euclidean distance measures in terms of accuracy. The KNN model with modified Canberra distance measure outperformed its original distance measure in terms of accuracy and recall and performed the same in terms of F score along with the Euclidean distance measure. The AUC score for both modified distance measures was 0.98. Thus, from the results, it can also be observed that as the size of the training data increases, the accuracy of the KNN model increases with our modified distance measures for the classification of the object.

(c) The KNN Model performance for the Diabetes Dataset

The KNN model has been trained for the diabetes dataset with 65% training size and 80% training size of the dataset. The performance measures for 65% of the training size for the KNN model with the difference distance measures are listed in Table 5.

Table 5: Performance Measures of the KNN model for the Diabetes dataset for 65% training size.

Distance Measure	Accuracy	Precision	Recall	F-Score
Modified Sorensen	0.717	0.690	0.690	0.690
Modified Canberra	0.717	0.690	0.690	0.690
Sorensen	0.725	0.700	0.700	0.700
Canberra	0.703	0.670	0.650	0.650
Euclidian	0.725	0.700	0.700	0.700

It can be seen from Table 5 that for 65% of the training size for the dataset, the KNN model with modified Canberra distance measures outperformed its original distance measure in terms of accuracy, precision, recall and F-score. The KNN model with modified Sorensen distance measure performed almost equally well as its original and Euclidean distance measures in terms of precision, recall and F-score. The AUC score for both modified distance measures was 0.76. The performance measures for 80% of the training size for the diabetes dataset for the KNN model with the difference distance measures are presented in Table 6.

Table 6: Performance Measures of the KNN model for the Diabetes dataset for 80% training size.

Distance Measure	Accuracy	Precision	Recall	F-Score
Modified Sorensen	0.708	0.670	0.660	0.660
Modified Canberra	0.708	0.690	0.690	0.690
Sorensen	0.734	0.700	0.700	0.700
Canberra	0.669	0.670	0.650	0.650
Euclidian	0.721	0.700	0.700	0.700

Table 6 shows that for 80% of the training data, the KNN model with modified Canberra distance measures outperformed its original distance measure in terms of accuracy, precision,

recall and F score and almost equally performed well as compared to the Euclidean distance measure. The AUC score for both modified distance measures was 0.71.

In addition to improved accuracy, precision, recall and F-score, the modified distance measures have several other advantages such it is easy to compute and implement. It is also applicable to small and large datasets of different domains for classification models. Therefore, the modified distance measures can be used in a variety of machine-learning applications because

(i) it can be used to improve the performance of classification algorithms by providing a more accurate measure of the similarity between data points.

(ii) It can also be used to improve the performance of clustering algorithms by grouping similar data points.

Thus, the modified distance measures are promising new methods for computing the similarity between data points. It has been shown to outperform existing distance measures in terms of accuracy, precision, recall and F-score.

4. Conclusion

This paper aims to compare the performance of the KNN model classification problem using Sorensen, Canberra and Euclidean distance measures with the modified Sorensen and Canberra distance measures. The proposes modified distance measures of Sorensen and Canberra which aims to overcome some of the limitations of existing methodology. The performance of the KNN models with modified distance measures is evaluated on several benchmark datasets and compared to its original Sorensen and Canberra along with Euclidean distance measures. In this study, the performance of the KNN classifier is evaluated on the basis of accuracy, precision, recall, F-score and AUC score using modified distance measures to improve the model classification accuracy and performance. Furthermore, we compared the performance of the KNN classifier with the modified distance measures and original distance measures including the Euclidean distance measure used for continuous data and found that the KNN model with modified distance measures outperformed in the case of small dataset iris and for large datasets such as breast cancer and diabetes datasets. It is also observed that as the size of training data increases, the performance measures increase for all the datasets. Therefore, the proposed modified distance measures may play a significant role in the classification problems of the datasets from the different domains. It would be useful for many classification techniques in machine learning.

References:

[1] Agrawal, R. & Ram, B. (2015). A Modified K-Nearest Neighbor Algorithm to Handle Uncertain Data. *2015 5th International Conference on IT Convergence and Security (ICITCS), Kuala Lumpur, Malaysia,* 1-4. DOI: 10.1109/ICITCS.2015.7292920.

[2] Alfeilat, H.A.A. Hassanat, A B.A., Lasassmeh, O., Tarawneh, A.S., Alhasanat, M.B.A., Salman, H.S.E. & Prasath, V.B.S. (2019). Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*. 7(4) 221-248. https://www.liebertpub.com/doi/10.1089/big.2018.0175

[3] Ali, N., Neagu, D. & Trundle, P. (2019). Evaluation of k-nearest neighbor classifier performance for heterogeneous data sets. *SN Applied Sciences.* 1, 1559 (2019). https://doi.org/10.1007/s42452-019-1356-9.

[4] Arya, S., & Mount, D. M. (1993). Approximate nearest neighbor queries in fixed dimensions. In Proceedings of the 4th Annual ACM/SIGACT-SIAM Symposium on Discrete Algorithms (SODA), 271–280.

[5] Bhatia, N. & Vandana (2010). Survey of Nearest Neighbor Techniques. *International Journal of Computer Science and Information Security*, 8 (2), 302-305.

[6] Blanco-Mallo, E., Morán-Fernández, L., Remeseiro, B., & Bolón-Canedo, V. (2023). Do all roads lead to Rome? Studying distance measures in the context of machine learning. *Pattern Recognition*, 141, 109646.

[7] Bramer, M. (2007). *Principles of data mining*. Springer.

[8] Chattopadhyay, N., Chattopadhyay, A., Gupta, S. S., & Kasper, M. (2019). Curse of dimensionality in adversarial examples. *In 2019 International Joint Conference on Neural Networks (IJCNN)*, 1-8.

[9] Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13 (1), 21-27.

[10] Deza, E. & Deza, M. M. (2009). *Encyclopaedia of distances*. Springer.

[11] Fix, E. & Hodges, J.L. (1951). Discriminatory analysis. Nonparametric discrimination; consistency properties. *Technical Report 4, USAF School of Aviation Medicine, Randolph Field, TX, USA*.

[12] Greche, L., Jazouli, M., Es-Sbai, N., Majda, A. & Zarghili, A. (2017). Comparison between Euclidean and Manhattan distance measure for facial expressions classification. *International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco*, 1-4, DOI: 10.1109/WITS.2017.7934618.

[13] Kamble, V.H. & Dale, M.P. (2022). Machine learning approach for longitudinal face recognition of children. *In Machine Learning for Biometrics*, 1-27.

[14] Lance, G.N. & Williams, W.T. (1967). Mixed-data classificatory programs I-Agglomerative systems. *Australian Computer Journal*, 1(1), 15-20.

[15] Mahesh, B. (2019). Machine learning algorithms - a review. International Journal of Science and Research. 9, 381–386.

[16] Manne, S., Kotha, S. & Fatima, S.S. (2012). Text categorization with K-nearest neighbor approach. *In Proceedings of the International Conference on Information Systems Design and Intelligent Applications* (INDIA) held in Visakhapatnam, India, 132, 413-420.

[17] Navin, J.R.N. & Pankaja, R. (2016). Performance analysis of text classification algorithms using confusion matrix. *International Journal of Engineering and Technical Research (IJETR)*, *6*(4), 75-78.

[18] Pandit, S. & Gupta, S. (2011). A Comparative Study on Distance Measuring Approaches for Clustering. *International Journal of Research in Computer Science*, 2(1), 29-31. DOI:10.7815/ijorcs.21.2011.011

[19] Prasatha, V.S., Alfeilate, H.A.A., Hassanate, A.B., Lasassmehe, O., Tarawnehf, A.S., Alhasanatg, M.B., & Salmane, H.S.E. (2017). Effects of distance measure choice on knn classifier performance-a review. p. 39, arXiv Preprint arXiv:1708.04321v3.

[20] Short, R. & Fukunaga, K. (1981). The optimal distance measure for nearest neighbor classification," in *IEEE Transactions on Information Theory*, 27 (5), 622-627, DOI: 10.1109/TIT.1981.1056403.

[21] Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, 5,1-34.

[22] Uddin, S., Haque, I., Lu, H., Moni, M.A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbor (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, 12(1), 1-11.https://doi.org/10.1038/s41598-022-10358-x

[23] Wettschereck, D. (1994). *A study of distance-based machine learning algorithms*. Oregon State University, Corvallis.

[24] Williams, W.T., & Lance, G.N. (1966). Computer programs for hierarchical polythetic classification ("similarity analyses"). *The Computer Journal*, 9 (1), 60-64.

[25] Xiubo, G., Tie-Yan, L., Qin, T., Andrew, A., Li, H., & Shum, H.Y. (2008). Query-dependent ranking using k-nearest neighbor. *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval,* 115-122.

[26] Xu, S., & Wu, Y. (2008). An algorithm for remote sensing image classification based on artificial immune B-cell network. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37, 107-112.

[27] Yang, Y., Ault, T., Pierce, T., & Lattimer, C.W. (2000). Improving text categorization methods for event tracking. *In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 65-72.