Transactions on Engineering and Computing Sciences - Vol. 11, No. 3
Publication Date: June 25, 2023
DOI:10.14738/tecs.113.14768.
Daulay, R. S. A., Efendi, S., & Suherman. (2023). Review of Literature on Improving the KNN Algorithm. Transactions on Engineering and Computing Sciences, 11(3). 63-72.

Review of Literature on Improving the KNN Algorithm

Raja Sakti Arief Daulay
Faculty of Computer Science and Information Technology,
Universitas Sumatera Utara, Indonesia

Syahril Efendi
Faculty of Computer Science and Information Technology,
Universitas Sumatera Utara, Indonesia

Suherman
Electrical Engineering Department,
Universitas Sumatera Utara, Indonesia

ABSTRACT

K-Nearest Neighbors (KNN) is a classification algorithm that has been widely used in the world of machine learning. The KNN algorithm classifies objects with learning data that are closest to the object. The special case where the classification is predicted based on the closest learning data is called the Nearest Neighbors algorithm. Classification is an important issue in processing big data, data science, and machine learning. KNN is one of the oldest, simplest, and also accurate algorithms for pattern classification and regression models. Many researchers have also improvised KNNs and the results obtained by these studies have changed a lot in terms of the accuracy of the results. The purpose of this paper is to see the improvement of several types of development of the KNN algorithm

Keywords: K-Nearest Neighbors, machine learning, improve algorithm, classification method, data mining

INTRODUCTION

The KNN algorithm has been widely used for research related to machine learning. Various roles are used for the classification method performed by KNN. Therefore, many studies have modified or improved the performance of the KNN. Starting with improving the accuracy of the classification, and combining KNN with the algorithm. Classification is an important issue in processing big data, data science, and machine learning. KNN is one of the oldest, simplest, and also accurate algorithms for pattern classification and regression models. KNN proposed in 1951 by Evelyn and Hodges and later modified by Cover and Hart, KNN has been identified as one of the top ten methods in data mining [1]. As a result, KNN has been studied over the last few decades and is widely applied in various fields. Thus, KNN comprises basic classifiers in many pattern classification problems such as pattern recognition, text categorization, ranking models, object recognition, and event recognition applications [2]. KNN is a nonparametric algorithm. Nonparametric means no parameters or a fixed number of parameters regardless of data size. Instead, the parameters will be determined by the size of the training data set, although no assumptions need to be made as to the distribution of the underlying data. Thus, KNN can be the best choice for any classification study that involves little or no prior knowledge of data distribution. In addition, KNN is one of the laziest learning methods [3]. This means that the storage of all training data will wait until the test data is generated, without having to create a learning model. Lag is not the only problem associated with the KNN classifier, apart from choosing the best K-Neighbors problem, choosing the best distance/similarity measure is an important issue, this is because the performance of the KNN classifier depends on the distance/similarity measure used [4][5].

Weaknesses in the KNN algorithm are used as material for researchers to try to improve the performance of KNN. So, there are many types of improvements or modifications to the KNN algorithm. Many KNNs are also combined with other machine learning techniques so that they can reduce the weaknesses of these KNNs. Therefore, this research will focus on introducing several modifications and improvements to the KNN algorithm so that it is hoped that it can become a guideline for future research.

RESEARCH METHODS

This research is a literature review of several articles related to the KNN algorithm. The review was carried out from several recent research efforts that utilize the KNN algorithm. After that, this study comes from several literacies and includes efforts to improve and modify the KNN algorithm which is divided into each research title.

The data collection process used to examine some of the literature is very useful for finding and obtaining research sources based on previous relevant research. Supporting theories, data, and information as references in the documentation.

RESULTS AND DISCUSSION

Metadata Information in the Web Pages and Expansion of the Query

The results of the review show that several studies have used the enhanced KNN algorithm to solve problems according to the needs of their respective research fields. An explanation of the improvement of the KNN algorithm and solving the problem in detail is shown below:

Effects of Distance Measure Choice on K-Nearest Neighbors Classifier Performance:
A Review — Analyze the performance of the KNN classifier using 11 distance measures. Including Euclidean Distance (ED), Manhattan Distance (MD), Minkowski Distance, Chebyshev Distance, Cosine Distance (CosD), Correlation Distance (CorD), Hamming Distance (HamD), Jaccard Distance (JacD), Standardized Euclidean Distance, and Spearman Distance. This research has been applied to eight binary synthetic data sets with various types of distributions generated using MATLAB. They divided each data set into 70% for the training set and 30% for the test set. The results show that the MD, Minkowski Distance, Chebyshev Distance, ED, Mahala Nobis Distance, and Standardized Euclidean Distance measures achieve the same accuracy and outperform the other tested distances. The data set was normalized before experimenting. To evaluate the performance of the KNN, measurements of accuracy, sensitivity, and specificity were calculated for each distance. The reported results show that using MD outperformed the other tested distances, with an accuracy rate of 97.8%, a sensitivity rate of 96.76%, and a specificity rate of 98.35%. No optimal distance metric can be used for all types of data sets

because the results show that each data set resembles a certain distance metric. The performance (measured by accuracy, precision, and recall) of KNN only drops by 20% while the noise level reaches 90%, this applies to all distances used. This means that a KNN classifier using one of the top 10 distances tolerates noise to a certain degree [5].

A new locally adaptive k-Nearest Neighbors algorithm based on discrimination class:
The Nearest Neighbors selection method from the KNN algorithm is not reliable enough. One reason is to ignore the influence of the spatial distribution of queries and training instances on classification. Therefore, a new concept of nearest centroid neighbor (NCN) is proposed, which considers the distribution of training examples in the query environment. A dependent Nearest Neighbors (dNN) algorithm that considers not only distance-determined similarity but also angle-defined dependency when selecting the nearest dependent neighbor of the query. Another reason is that the Nearest Neighbors selection method of the KNN algorithm only uses one-sided similarity from the query point of view. To increase the similarity between the query and the Nearest Neighbors, the neighbor information on the training examples has been taken into account in many jobs. The KNN algorithm only uses the Neighbors of the majority class in the k-neighborhood of the query to determine the classification result of the query and completely ignores information including Nearest Neighbors in other classes which can lead to poor classification results when the number of Nearest Neighbors in different classes is not much different. In addition, the KNN algorithm uses a single, fixed k value for all queries, but this k cannot be optimal for queries at different spatial locations [6].

Study and Observation of the Variation of Accuracies:
of KNN, SVM, LMNN, and ENN Algorithms on Eleven Different Datasets from UCI Machine Learning Repository — The KNN and SVM algorithms are similar in both cases in that certain areas are considered to find the most probabilities. The enhanced KNN shows better accuracy on web text location in groups than the automated KNN algorithm. The main benefit of SVM is the ability to handle large data sets which KNN fails to do in some cases. On the other hand, the weakness is the high cost and complexity. To overcome the deficiencies in KNN, another method named Extended Nearest Neighbors (ENN) was proposed and its performance was observed which implies that ENN shows better performance than traditional KNN [7]. At ENN, apart from considering the closest data test sample, it is also calculated in the closest data which provides better accuracy. The four machine learning algorithms, namely KNN, ENN, SVM, and LMNN, have shown a strong impact on data classification in various sectors. Even though currently KNN and SVM are replaced by ENN and LMNN, because they are more perfect and accurate, the vital role of KNN and SVM cannot be ignored [8].

DDoS Attack Detection Method Based on Improved KNN with the Degree of DDoS Attack in Software:
Defined Networks — Distributed Denial of Service (DDoS) attacks have disrupted network availability for decades and there is still no effective defense mechanism against them. In this research, they propose two methods to detect DDoS attacks on SDN. One method adopts DDoS attack rates to identify DDoS attacks. Another method uses the enhanced K-Nearest Neighbors (KNN) algorithm based on Machine Learning to find DDoS attacks. Combines Support Vector Machine (SVM) classification algorithms to build DDoS attack models [9]. In their method, six

feature values are introduced. Their experimental results show a low false alarm rate for TCP and UDP traffic, but a high false alarm rate for ICMP traffic. Propose a mechanism using the Cognitive-Inspired Computing (CIC) classification algorithm and SVM to detect DDoS attacks. Meanwhile, the detection accuracy still needs to be improved. Based on the concept of this research, a detection algorithm called an Algorithm based on the Degree of Attack (DDADA) is proposed. In addition to further increasing detection efficiency, another detection algorithm called the DDoS Detection Algorithm based on Machine Learning (DDAML) was introduced to identify DDoS attacks. The experimental results show that the proposed algorithm can better identify DDoS attacks and has achieved a higher detection rate compared to existing solutions. The experimental results show that the DDAML algorithm can outperform other algorithms on different performance measurements.

Classification Method of Teaching Resources Based on Improved KNN Algorithm:
Manufacturing personnel performance modeling based on the improved KNN algorithm. Since then, many experts and researchers have achieved certain research results in this field of technology, such as the well-known intelligence scientists Spark and Salton. Since the 1980s, traditional knowledge engineering techniques have been applied in this field. According to the knowledge provided by experts, rules are formed and classifiers are created manually. This is a good classification in several corpora. However, when dealing with large-scale data sets, the methods are limited. The text preprocessing process is enhanced by the proposed strategy using the KNN algorithm. The aim of the traditional KNN algorithm problem in classifying teaching resources in primary and secondary schools is to propose a good KNN algorithm based on a density tailoring scheme. A sample of high-density regions in the sample space is cut before the KNN algorithm is run. The problem of misclassification caused by the uneven distribution of the spatial density of samples is solved. At the same time, the time complexity of KNN classification is reduced. The corresponding K and input parameters were determined by comparison experiments and the effectiveness of the enhanced algorithm was verified. The results showed that the method of classifying data sources for elementary and junior high schools based on the KNN algorithm was feasible and effective [10].

Fast-Density Peak Clustering for Large-Scale Data Based on KNN:
A vantage-point tree or VP tree is very similar to a K-D tree in that it separates data in metric space by choosing a position in space (viewpoints) and dividing the data points into two partitions: those that are closer to the vantage point than the threshold and those that are not. By repeatedly applying this procedure to partition the data into smaller and smaller sets, a data tree structure is created in which Neighbors in the tree tend to be Neighbors in space. FLANN is a library for performing fast approximate Nearest Neighbors in high dimensional spaces. FLANN contains a collection of algorithms for Nearest Neighbors search and a system for automatically selecting the best algorithm and optimal parameters depending on the data set. Using a cover tree to accept the KNN for each point and the proposed KNN density replaces the original density defined in Dpeak, which results in a major improvement for density calculations [11].

Cost-Sensitive KNN Classification:

KNN method for classifying big data [12]. The first KNN performs k-means clustering to separate the entire dataset into several parts. Then each subset is classified by the KNN method. k_{Tree} and k*Tree use different numbers of Nearest Neighbors for KNN classification. The k_{Tree} method requires less operational cost but achieves a similar classification accuracy compared to the KNN method which assigns different K values to different test samples. The k*Tree method is an extension of kTree. Namely speeding up the testing phase by storing training sample information in the kTree leaf nodes, such as the training samples located in their KNN leaf nodes and the closest Neighbors of the KNN. Create a KNN using only a subset of the training samples in the leaf nodes. This paper presents two approaches, namely the direct-cs-KNN classifier and the distance-cs-KNN classifier which aim to make the KNN classification cost-sensitive to minimize the cost of misclassification. For the efficiency of several useful methods including smoothing the k arrangements at minimum cost, feature selection of CS and CS stacking are significantly coupled to the CS-KNN classifier [13].

Machine Learning Methods for Cyber Security Intrusion Detection:

Datasets and comparative study — In this study [14], researchers used the RNN deep learning (recurrent neural network) method, deep neural networks (DNN), restricted Boltzmann machines (RBM), deep belief networks (DBN), convoluted neural networks (CNN), deep Boltzmann machines (DBM), and deep autoencoders (DA) have implemented it in the CSE-CIC-IDS2018 and Bot-lot datasets. Then the classification success of deep learning and the classification time of these data sets compared to deep learning methods are examined and, in this sense, the 35 attack detection data sets used in the literature are divided into categories. CSE-CIC-IDS-2017 dataset, because the existing dataset does not meet current intrusion detection needs. A testing environment consisting of network attackers and victims has been prepared to create a data set. In testing the attack environment such brute force, Heartbleed attacks, botnets, DOS, DDoS, web attacks, and infiltration attacks are regulated. In addition, system performance is evaluated using machine learning methods. With the development of smart technology, the Internet is used in every area of daily life. With the widespread use of the internet, the types of attacks are growing day by day. To prevent attacks, these attacks must be detected first. In this case, an IDS system has been developed to detect attack traffic and an IDS data set has been created to simulate the type of attack [15], [16].

Internet Digital Economy Development Forecast Based on Artificial Intelligence and SVM-KNN Network Detection:

Several studies have proposed a KNN clustering model based on density peaks that can detect attacks more effectively and introduce density to KNNs. The KNN process does not require many parameters and the iterative process is based on density. The literature proposes a new Verves intrusion detection model in which a new data mining classification method is incorporated i.e., as an alternative to Verves nets using the algorithm. Based on literature research related to the SVM-KNN classification algorithm and the KNN classification algorithm, according to the classification characteristics of the current KNN classification algorithm and the combination of advantages of the harmonic-weighted KNN algorithm and the support vector machine algorithm, the combination of the SVM and KNN algorithms will be combined to build an algorithm SVM-based harmonic-weighted KNN to improve the classification performance of data sets [17].

Comparison and Analysis of Logistic Regression, Naive Bayes, and KNN Machine Learning Algorithms for Credit Card Fraud Detection:
Researching and examining the presentation of the Decision Tree, Random Forest, SVM, and Logistic Regression classifier algorithms. Methods used on raw and pre-treated information. From the investigations that have been carried out, the results show that Logistic Regression has an accuracy of 97.7% while SVM shows an accuracy of 97.5% and a decision tree shows an accuracy of 95.5% but the best results are obtained with a random forest with an exact precision of 98.6%. The results obtained are due to the reason that the random forest shows the most precise and high accuracy of 98.6% in the credit card fraud detection problem with the dataset provided by ULB. Examine various machine learning classifiers trained on public datasets to analyze the correlation of certain factors with fraud. Better metrics are used to determine false negative rates and random sampling performance is measured to address class imbalance of the data set [18]. SVM performs better in detecting credit card fraud under realistic conditions. Comparisons between deep learning and regression algorithm models were performed to determine which algorithm and combination of factors provide the most accurate method for classifying credit card transactions as fraudulent or non-fraudulent. The best algorithm for analyzing data sets with a ratio of fraudulent and non-fraudulent transactions close to 1:1 is the Random Forest Classifier, assuming the distribution of fraud to non-fraud of the test and training sets is the same [19]– [21]. Logistic Regression (LR) shows optimal performance for all data proportions compared to Naïve Bayes (NB) and K-Nearest Neighbors (KNN). LR managed to get higher accuracy compared to Naïve Bayes and KNN. LR shows maximum accuracy of 95%, NB shows 91%, and KNN 75%. The LR technique also shows better sensitivity, Specificity, Precision, and F-Measure compared to the NB and KNN techniques. It has also been observed that a supervised technique (LR and Naïve Bayes) shows better results in every case compared to an unsupervised KNN technique.

An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN:
A distance-based KNN classifier of comprehensive average through multi-comprehensive average cost as well as a structured comprehensive-average stretch based on representative comprehensive average was proposed in 2019. In the proposed method, multi-local vector averaging of samples The query given in each class is calculated by adopting the class-specific K-Nearest Neighbors. By obtaining k local averaged vectors per class, k generalized spaced averages are determined and used in designing a defined nested comprehensive average stretch. A new ensemble technique for the KNN algorithm was also proposed in 2019. A multimodal noise-based ensemble algorithm called Reduced Random Subspacebases Bagging (RRSB) was proposed which yields accurate yet diverse component classifiers to improve ensemble classification performance. For k values, the proposed RRSB system appears to be more robust than the other techniques. In this proposed technique, enhanced speech emotion recognition is performed on the six basic emotions anger, joy, sadness, neutral, surprise, and fear. As an advanced research methodology, pre-processing is performed using Pattern Recognition Neural Network (PRNN) and KNN algorithms while feature extraction is performed using a multilevel structure consisting of Mel Frequency Cepstral Coefficient (MFCC)

and Gray Level Co-Occurrence Matric (GLCM). The results obtained were compared for accuracy, level of precision, and f Measure with standard algorithms such as the Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) and were recognized as better output than standard algorithms [22].

A Novel Weighted KNN Algorithm Based on RSS Similarity and Position Distance for Wi-Fi Fingerprint Positioning:
Propose a Weighted KNN (WKNN) method to assign different weights by defining the correlation coefficient between Access Points and achieve room-level positioning accuracy. Propose cluster-filtered WKNN algorithm. The affinity propagation clustering algorithm groups the closest Reference Points (RP) according to the distance of their positions from each other, and outliers are filtered to order the subset with the greater number of RPs. RP with this method is not based on the position distance between the RP and the Test Point (TP). Therefore, some RPs close to TP can be discarded as outliers, resulting in large positioning errors. By combining the weighted Euclidean distance with the known position information from the RP, position estimates are designed and used to improve the algorithm. The average positional accuracy of the proposed algorithm outperforms Euclidean-WKNN by 45.28%, 38.41%, and 20.60%, outperforms the SVM-based algorithm by 40.36%, 37.94%, and 33, respectively. 74% across three databases [23].

A Novel Ensemble Method for k-Nearest Neighbors:
Many research papers have investigated KNN ensembles intending to improve their performance. For example, this study [24] applies several distance metrics to produce diverse ensemble members, where distance metrics are needed as a learning parameter distraction. Ishii's [25] research combines KNN by using a Genetic Algorithm (GA) to weigh different distance functions. Several random subspaces are used to derive the component KNN classification investigated by Ho who trains each KNN on a random attribute subset rather than the entire attribute space. In this study, a new multimodal perturbation-based ensemble, reduced random subspace-based bagging (RRSB) algorithm is proposed which produces an accurate yet diverse component classifier to improve ensemble classification performance. Experimental results from several UCI data sets indicate that the proposed method can improve classification performance in most cases. Compared to other methods, the RRSB is strong with different k values.

Explaining and Improving Model Behavior with k Nearest Neighbors Representations:
The KNN language model (KNN-LM) has been introduced, extending existing pre-trained language models by interpolating subsequent word distributions with the KNN model. The KNN-LM combination is a met-learner that can be effectively tuned to memorize and retrieve rare long-tail patterns [26]. This work adapts KNN-LM to text classification by using input examples as a context like implementing deep-KNN via neural network representations for image classification. This study found that the KNN for each test sample can provide useful information about how the model makes its classification decisions, to misclassify the sample. From the analysis, it is also observed that the KNN of misclassified test samples often shows mislabeled samples, which gives this approach an application for correcting mislabeled examples in training sets. By finding the most common Nearest Neighbors across the test set,

they can also identify the subset of highly influential training examples and obtain corpus-level interpretations of model performance [26]. Finally, it examines the utility of falling back to KNN-based classification when model confidence is low. KNN-learned decision boundary analysis of hidden representations suggests that KNN learns fine-grained decision boundaries that can help make it more robust to small changes in the text that lead to basic truth labels, but not inverted model predictions.

Voice Disorder Identification by using Hilbert-Huang Transform (HHT) and K Nearest Neighbors (KNN):

When determining the classification method, KNN is a sample classification according to the closest sample category. Compared to other methods, KNN is more suitable for sample pool partitioning where class domains intersect or overlap [27]. Its training complexity is lower than SVM. In addition, compared to algorithms such as Naïve Bayes, KNN has no assumptions about the data, high accuracy, and is not sensitive to abnormal points. KNN is used in many fields. Xinyu Li [28] adds the KNN algorithm to the machine vision system to realize real-time monitoring of small faults. The automatic detection of speech disturbances has a very important clinical significance for diagnosing dysphonia [27], [29]. This paper presents an automatic classification method for voice and normal speech disturbances based on HHT and KNN. Experimental results show that this method has good performance, 93.3% accuracy, 93% precision, 95% recall rate, 94% F1-score, and 97.6% AUC [27]. The experimental results verify the validity and reliability of the feature extraction method. Experiments show that the KNN model not only obtains good classification results but also improves the generalization performance of the classifier.

With the improvement of the KNN algorithm, it is hoped that it will not rule out the possibility of only limited knowledge data but can develop more variants of KNN. Not all techniques can be developed for KNN but the table above is proof that existing algorithms can be developed so that they become up-to-date knowledge. The KNN algorithm continues to be developed by several researchers.

CONCLUSION

The KNN algorithm is a supervised learning algorithm that is used to perform classification or regression based on the Nearest Neighbors of new data. From the table above, it can be observed that several techniques can be used to improve the performance of the KNN algorithm, including:

1. Data standardization: Before entering data into the KNN algorithm, it is important to standardize the data. This is done to ensure that each variable has the same scale and corrects abnormal data distribution.
2. Use of Weights: Provides a different weight for each neighbor based on the distance between the data and the point you want to predict. The closer the distance, the greater the weight. This can improve prediction accuracy.
3. Feature selection: Select the most relevant features that have a large influence on the prediction results. This can help reduce data dimensions and speed up computation time.

4. Cross-validation: This technique can help avoid overfitting and underfitting in the KNN model. Cross-validation can help evaluate model performance on previously unseen data.
5. Use appropriate distance metrics: Choose the appropriate distance metrics for the problem you want to solve. Some examples of commonly used distance metrics are Euclidean, Manhattan, and Cosine distance.

By applying these techniques, we can improve the performance of the KNN algorithm and obtain more accurate predictions.

References

[1]. L. Wang, "Research and Implementation of Machine Learning Classifier Based on KNN," in IOP Conference Series: Materials Science and Engineering, 2019.

[2]. A. Mehta, D. Goyal, A. Choudhary, B. S. Pabla, and S. Belghith, "Machine Learning-Based Fault Diagnosis of Self-Aligning Bearings for Rotating Machinery Using Infrared Thermography," Math. Probl. Eng., 2021.

[3]. P. Tamrakar and S. P. Syed Ibrahim, "Lazy Learning Associative Classification with WKNN and DWKNN Algorithm," ITM Web Conf., 2021.

[4]. S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-Nearest Neighbors (KNN) algorithm and its different variants for disease prediction," Sci. Rep., 2022.

[5]. H. A. Abu Alfeilat et al., "Effects of Distance Measure Choice on K-Nearest Neighbors Classifier Performance: A Review," Big Data, vol. 7, no. 4, pp. 221–248, 2019.

[6]. Z. Pan, Y. Wang, and Y. Pan, "A new locally adaptive k-Nearest Neighbors algorithm based on discrimination class," Knowledge-Based Syst., vol. 204, p. 106185, 2020.

[7]. J. Na, Z. Wang, S. Lv, and Z. Xu, "An Extended K Nearest Neighbors-Based Classifier for Epilepsy Diagnosis," IEEE Access, 2021.

[8]. M. M. R. Khan, R. B. Arif, A. B. Siddique, and M. R. Oishe, "Study and observation of the variation of accuracies of KNN, SVM, LMNN, ENN algorithms on eleven different datasets from UCI machine learning repository," 4th Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2018, pp. 124–129, 2018.

[9]. S. Dong and M. Sarem, "DDoS Attack Detection Method Based on Improved KNN with the Degree of DDoS Attack in Software-Defined Networks," IEEE Access, vol. 8, pp. 5039–5048, 2020.

[10]. Y. An, M. Xu, and C. Shen, "Classification method of teaching resources based on improved KNN algorithm," Int. J. Emerg. Technol. Learn., vol. 14, no. 4, pp. 73–88, 2019.

[11]. Y. Chen et al., "Fast density peak clustering for large scale data based on KNN," Knowledge-Based Syst., vol. 187, no. xxxx, 2020.

[12]. C. Ma and Y. Chi, "KNN Normalized Optimization and Platform Tuning Based on Hadoop," IEEE Access, 2022.

[13]. S. Zhang, "Cost-sensitive KNN classification," Neurocomputing, vol. 391, pp. 234–242, 2020.

[14]. I. F. Kilincer, F. Ertam, and A. Sengur, "Machine learning methods for cyber security intrusion detection: Datasets and comparative study," Comput. Networks, vol. 188, no. October 2020, p. 107840, 2021.

[15]. S. Al-E'mari, M. Anbar, Y. Sanjalawe, S. Manickam, and I. Hasbullah, "Intrusion detection systems using blockchain technology: A review, issues and challenges," Computer Systems Science and Engineering. 2022.

[16]. A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," Cybersecurity, 2019.

[17]. J. Fu, X. Zhou, and G. Mei, "Internet Digital Economy Development Forecast Based on Artificial Intelligence and SVM-KNN Network Detection," Comput. Intell. Neurosci., vol. 2022, 2022.

[18]. F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," Int. J. Inf. Technol., vol. 13, no. 4, pp. 1503–1511, 2021.

[19]. T. Zhou et al., "MSIFinder: a python package for detecting MSI status using random forest classifier," BMC Bioinformatics, 2021.

[20]. N. H. Z. Abidin et al., "Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier," Int. J. Adv. Comput. Sci. Appl., 2020.

[21]. S. Hamsa, Y. Iraqi, I. Shahin, and N. Werghi, "An Enhanced Emotion Recognition Algorithm Using Pitch Correlogram, Deep Sparse Matrix Representation and Random Forest Classifier," IEEE Access, 2021.

[22]. J. Umamaheswari and A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Perspectives Prospect. Com. 2019, pp. 177–183, 2019.

[23]. B. Wang et al., "A Novel Weighted KNN Algorithm Based on RSS Similarity and Position Distance for Wi-Fi Fingerprint Positioning," IEEE Access, vol. 8, pp. 30591–30602, 2020.

[24]. Y. Zhang, G. Cao, B. Wang, and X. Li, "A novel ensemble method for k-Nearest Neighbors," Pattern Recognit., vol. 85, pp. 13–25, 2019.

[25]. N. Ishii, E. Tsuchiya, Y. Bao, and N. Yamaguchi, "Combining classification improvements by ensemble processing," in Proceedings - Third ACIS International Conference on Software Engineering Research, Management, and Applications, SERA 2005, 2005.

[26]. N. F. Rajani, B. Krause, W. Yin, T. Niu, R. Socher, and C. Xiong, "Explaining and Improving Model Behavior with k Nearest Neighbors Representations," 2020.

[27]. L. Chen, C. Wang, J. Chen, Z. Xiang, and X. Hu, "Voice Disorder Identification by using Hilbert-Huang Transform (HHT) and K Nearest Neighbors (KNN)," J. Voice, vol. 35, no. 6, pp. 932.e1-932.e11, 2021.

[28]. X. Li, T. Qiao, Y. Pang, H. Zhang, and G. Yan, "A new machine vision real-time detection system for liquid impurities based on dynamic morphological characteristic analysis and machine learning," Meas. J. Int. Meas. Confed., 2018.

[29]. G. Tartarisco et al., "Artificial Intelligence for Dysarthria Assessment in Children with Ataxia: A Hierarchical Approach," IEEE Access, 2021.