IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 1, March 2024, pp. 1149~1159
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i1.pp1149-1159

K-centroid convergence clustering identification in one-label per type for disease prediction

Minh Long Hoang, Nicola Delmonte
Department of Engineering and Architecture, University of Parma, Parma, Italy

Article Info

Article history:

Received Jul 13, 2023
Revised Sep 25. 2023
Accepted Sep 30, 2023

Keywords:

Disease prediction
K-centroid-convergence clustering identification
Machine learning
Medical science
Semi-K clustering

ABSTRACT

Disease prediction is a high demand field which requires significant support from machine learning (ML) to enhance the result efficiency. The research works on application of K-means clustering supervised classification in disease prediction where each class only has one labeled data. The K-centroid convergence clustering identification (KC^3I) system is based on semi-K-means clustering but only requires single labeled data per class for the training process with the training dataset to update the centroid. The KC^3I model also includes a dictionary box to index all the input centroids before and after the updating process. Each centroid matches with a corresponding label inside this box. After the training process, each time the input features arrive, the trained centroid will put them to its cluster depending on the Euclidean distance, then convert them into the specific class name, which is coherent to that centroid index. Two validation stages were carried out and accomplished the expectation in terms of precision, recall, F1-score, and absolute accuracy. The last part demonstrates the possibility of feature reduction by selecting the most crucial feature with the extra tree classifier method. Total data are fed into the KC^3I system with the most important features and remain the same accuracy.

Corresponding Author:

Minh Long Hoang
Department of Engineering and Architecture, University of Parma
Parma 43124, Italy
Email: minhlong.hoang@unipr.it

1. INTRODUCTION

Nowadays, machine learning (ML) development has contributed a significant role to the classification field in medical science with ever-growing clinical datasets [1]-[6]. ML is particularly valuable in the healthcare industry because it can interconnect massive amounts of data produced daily within electronic health records. Furthermore, healthcare providers can generate more predictive methodology, forming a more unified system with enhanced care delivery and patient-based processes, especially in disease prediction. However, to make the ML work properly, the most necessary part is to collect a good amount of data, which may become challenging in many situations when the labeled data is really limited. When there is only a single labeled data per disease class for training, it is still a huge challenge for ML. This work proposes a K-centroid convergence clustering identification (KC^3I) system, which is able to work with this particular case and achieve high accuracy of disease prediction. This approach is highly effective not only for disease prediction and also other applications which require ML support when the labelled data is limited. In addition, this research presents a method to reduce the input features but still guarantees the same accuracy for ML model.

This work mainly studies the harsh case where each class only has one labeled data. For supervised learning (S.L.) [7]-[9], it requires a large number of training features labeled with coherent outputs [10].

Consequently, the S.L. cannot operate properly in this specific case. The Unsupervised Learning (U.L.) like clustering technique [11], [12] can work without the labeled data, which is mostly based on an objective function of similarity or dissimilarity measures where data set can be represented by finite cluster prototypes with their own objective functions [13].

Generally, the K-Means algorithm is the most known and used clustering method [14]. K-means clustering has been widely researched with various extensions in the literature and employed in various substantive areas [15]-[18]. Another traditional U.L. clustering is hierarchical clustering which groups data points into a series of clusters in a tree-like structure and visually represented in a hierarchical tree called a dendrogram. However, they are usually used to analyze the relationship between the input features or detect significant patterns [19], not to predict the specific name of the output class, especially in the case of multiple categories [20].

The combination between U.L. and S.L. can be the right approach for this circumstance, but despite that, semi-supervised methods [21]-[23] also involve a sufficient number of labeled data for type prediction. Among many algorithms, seeded k-means clustering supervised classification methods may work with only a small subset of labeled data [24]-[26]. Nevertheless, no research has been carried out about working with unique labeled data per cluster. Hence, this research depicts a system named K-centroid convergence clustering identification (KC^3I), which proceeds through the logical stages to predict specific names of the concerned categories. The KC^3I considers each labeled data as the initial centroid for the different cluster in the indexed array, then pulls all other neighbors to each centroid based on the minimum distance adopted distance measure to find the distance between any two vectors. After 1st iteration, all the data are converged into the specific cluster. Each cluster's mean is calculated and becomes the new centroid for the 2nd convergence with these updated centroids. An important factor here is the 'Dictionary box', which plays a role in recording all the indexed centroids from the beginning of the process to the end of the training state. It matches each index centroid with a related disease. After the training process, new input value of features that enters the KC^3I model will be evaluated to inject into an indexed cluster and converted to a corresponding disease type by dictionary box as demonstrated in Figure 1, Figure 2, and Figure 3.

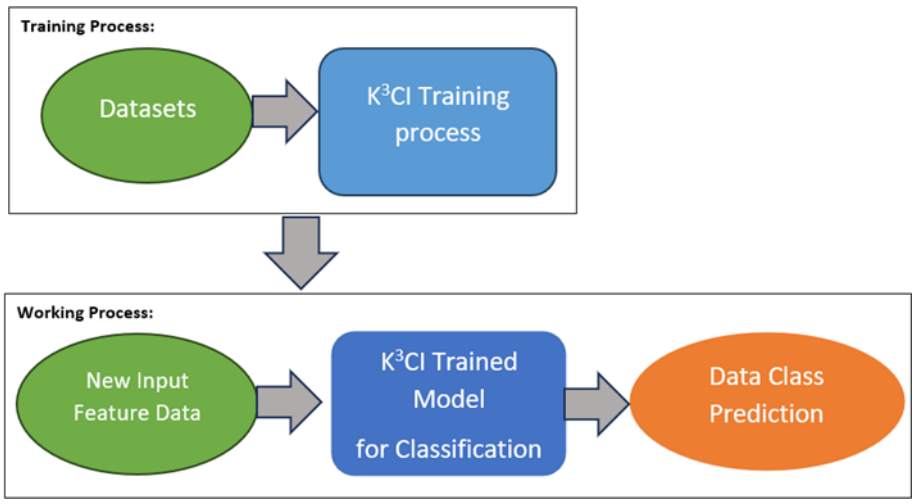

Figure 1. System architecture

The applied case is the disease prediction based on the input symptoms, with open access medical dataset [27], which is composed of 132 parameters of symptoms on which 41 different disease types. The dataset contains the actual label which are used to validate the accuracy of the proposed technique. During the training process, all true labels are neglected, there is only one labeled data for each disease type as the initial centroid for each cluster.

This application aims to achieve the absolute accuracy of prediction, respected with the true label because it may be critical if only one error in disease prediction for patients in healthcare. In addition, the feature reduction based on the extra-trees classifier [28], [29] will be examined by filtering out the most essential features in the KC^3I model but still guaranteeing the same prediction accuracy. Table 1 summarizes the main contribution of this paper to scientific development.

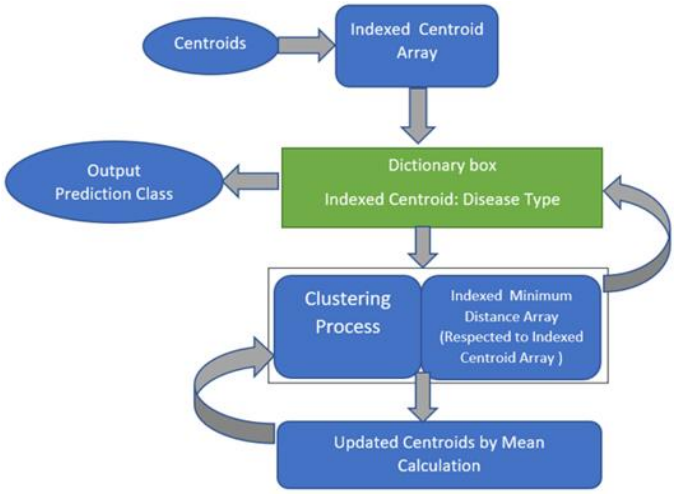

Figure 2. KC3I training chart

Table 1. Summary of research contribution

Research contribution
– New approach on K-means clustering supervised classification with only one labelled data per cluster.
– Classification process with Dictionary box for accurate cluster name between multiple groups
– Practical application to large dataset of disease prediction based on the symptoms with 100 % of accuracy.
– Detail validation description and result explanation of ML system performance.
– Most dominant features are filtered out by, which maintain the same result accuracy with Tree-Based Classifiers to decrease the required number of symptoms for disease prediction.

The paper is organized as follows: the 1st part is a brief description of the utilized dataset, then the training and operation of the KC^{3}I system will be described in detail. In the next section, two stages of validations are carried out, including training verification and test prediction. Finally, it is about the feature reduction part and conclusion.

2. METHOD

2.1. Utilized data

The medical dataset [27] is used for applying knowledge to medical science, and making physicians' tasks easy is this dataset's primary purpose. The dataset is made up of 132 parameters of symptoms on which 41 different disease types can be predicted. These symptoms are mapped to the corresponding diseases as true label.

**Note: In the website mentions about 42 diseases but the downloaded data only has 41 different diseases without repetition. Thus, the research was carried out with 41 disease types with their relative data.*

In this research, all of the true labels are neglected. Another file is created, containing only input features and its disease label for each type as the initial centroid for each cluster. Hence, this centroid file has 41 samples, which are behalf of 41 disease types.

All the input features are scaled with the standardization as (1),

$$x_{scaled} = \frac{x_l - mean}{Std} \quad (1)$$

Where: x_l is features data of sample l.

2.2. Training process

We consider the harsh case when it is difficult to collect the raw data label, which happens in many circumstances. There is only one labeled data for each category. In the KC^{3}I model, each labeled data will be the centroid of each cluster. They are put into an array following an arranged index, so each array index is behalf of one specific disease type. The centroid index and name of coherent disease are saved into the 'dictionary box'.

At the next stage, each centroid starts pulling other data to their cluster based on the Euclidean distance [30]. Each input feature is fed into the distance calculation with centroids. The sample will join the cluster with the minimum distance between the sample and that cluster centroid.

In the clustering identification stage, each calculated distance is put into an array index following the previous centroid index. The array index of minimum distance is detected, and then inserted into 'index box'. In the dictionary process, the model converts that index number into the corresponding disease type.

After 1st training iteration, all the data in each cluster will be averaged, and the new mean value will be the new centroid. The training iteration will occur 2 times to achieve the stable centroid then it is ready for new data prediction. In n-dimensional space where n is equivalent to the feature number:
Centroid C = [c_1, c_2, ..., c_n]; Corresponding disease = [Fungal infection, Allergy…, Impetigo].
With Sample X = [x_1, x_2, ..., x_n].
As demonstrated in Table 2, x_1 is the value of symtom1, such as itching has a value of 0 or 1.
x_n is the value of the symptom.

Table 2. Dictionary box content

Centroids and Their index	Corresponding diseases
C = [c_1, c_2, ..., c_n] Index = [0,1…,n]	[Fungal infection, Allergy, …, Impetigo]

The distance between a single coordinate can be calculated as (2):

$$dj = xj - cj \quad (2)$$

Where j is the element index

$$Euclidean\ distance = \sqrt{{d_1}^2 + {d_2}^2 + \cdots {d_n}^2} \quad (3)$$

2.3. Operation

After the training process, the KC^{3}I is able to predict the output disease based on the input features. As the clustering identification stage, new input samples with their features enter the model, and the Euclidean distances with each centroid are calculated and arranged into an array index, following the centroid index. The array index of minimum distance is detected in the 'Indexed Minimum Distance Array' stage, and then passed to 'Dictionary box' to accomplish the corresponding disease type. All the procedures are illustrated in Figure 3 chronologically.

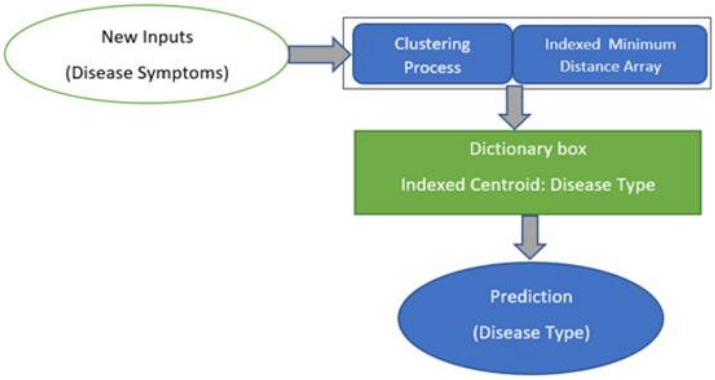

Figure 3. KC3I working chart

2.4. Validation metrics

To validate the KC^{3}I technique, the ML factors were calculated: precision, Recall, and F1-Score based on True Positive (T.P.A.), False Positive (F.P.A.), and False Negative (F.N.A.) of class A.

- T.P.A. is the number of predictions where the classifier correctly predicts class A.
- F.P.A. is the number of objects that do not belong to class a but are predicted as class A.
- F.N.A. is the number of objects from class A predicted to another class.
- Precision validates the number of the class predictions that actually belong to that class.

$$Precision = \frac{TP_A}{TP_A + FP_A} \quad (4)$$

– Recall indicates missed predictions of the class. In multiple classification, recall is determined as the true class number across all types divided by number of true positives and false negatives across all categories.

$$Recall = \frac{TP_A}{TP_A + FN_A} \quad (5)$$

– F1-Score provides a single score that balances the concerns of precision and recall in one number. Similar to precision and recall, a F-Measure score range is from 0.0 to 1.0.

$$F1\text{-}Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

– Macro average (avg) is a straightforward among the numerous averaging methods. The macro F1 score is the mean of all the per-class F1 scores. This method treats all classes equally regardless of their support values.

$$Marco\ avg = \frac{F1_1 + F1_2 + \cdots + F1_n}{n} \quad (7)$$

Where n is the class number in concern $F1_n$ is F1 score of class n.

– The weighted-average F1 score is calculated by taking the mean of all per-class F1 scores while considering each class's support. The 'weight' essentially refers to the proportion of each class's support relative to the sum of all support values.

$$\text{Weighted avg} = \frac{F1_1 * N_1 + F1_2 * N_2 + \cdots + F1_n * N_n}{n} \quad (8)$$

3. MODEL VALIDATION AND RESULT ANALYSIS

The dataset includes 4920 samples, divided into 70% as training data and 30% of the rest for testing. After the training process, the updated centroids are used in the validation process of the KC^{3}I system. The system was designed by Python, a high-level, general-purpose programming language [31], based on ML library of Scikit-learn [32]. The integration of Scikit-learn to Python has been utilized in many ML applications effectively [33]-[35]. The proposed system can potentially speed up the large amount of data from the sensors [36]-[37] in health monitoring.

As shown in Figure 4, the accuracy verification is divided into 2 stages:

– Stage 1: 3444 samples are used for the 1st stage. The trained dataset as input to see whether this semi-supervised technique can output the appropriate disease type.
– Stage 2: If the first validation shows an acceptable result, the test dataset will be fed into the system for further examination.

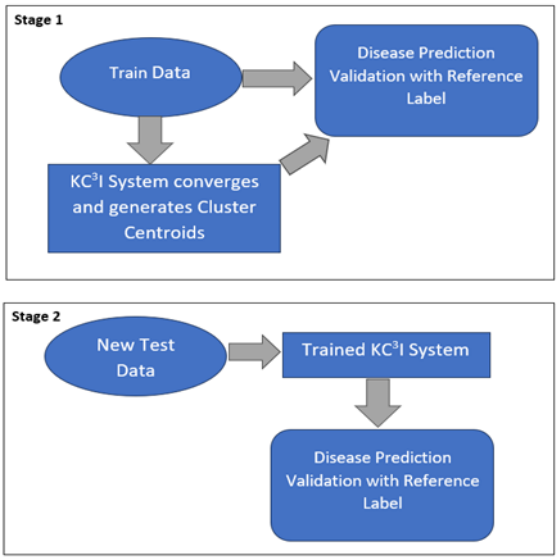

Figure 4. Two stages of system accuracy verification

3.1. Stage 1

After the training process, all the predictions are compared with the true label for the accuracy validation. Table 3 reports the impressive operation of the KC^{3}I technique since all the model predictions match the true label. As our demand, the model is supposed to make highest accuracy as possible because it is a huge issue if there is a small mistake in medical prediction. The system reaches the perfect performance in Precision and Recall so there is no misprediction between disease types. As a result, the training validation is sufficient for the next stage, about receiving the test data to output the prediction.

Table 3. Classification report of training data

	Disease type	Precision	Recall	F1-score	Support Number
Disease types	Paroxysmal Positional Vertigo	1	1	1	84
	AIDS	1	1	1	84
	Acne	1	1	1	84
	Alcoholic hepatitis	1	1	1	84
	Allergy	1	1	1	84
	Arthritis	1	1	1	84
	Bronchial Asthma	1	1	1	84
	Cervical spondylosis	1	1	1	84
	Chicken pox	1	1	1	84
	Chronic cholestasis	1	1	1	84
	Common Cold	1	1	1	84
	Dengue	1	1	1	84
	Diabetes	1	1	1	84
	Dimorphic hemorrhoids (piles)	1	1	1	84
	Drug Reaction	1	1	1	84
	Fungal infection	1	1	1	84
	GERD	1	1	1	84
	Gastroenteritis	1	1	1	84
	Heart attack	1	1	1	84
	Hepatitis B	1	1	1	84
	Hepatitis C	1	1	1	84
	Hepatitis D	1	1	1	84
	Hepatitis E	1	1	1	84
	Hypertension	1	1	1	84
	Hyperthyroidism	1	1	1	84
	Hypoglycemia	1	1	1	84
	Hypothyroidism	1	1	1	84
	Impetigo	1	1	1	83
	Jaundice	1	1	1	84
	Malaria	1	1	1	84
	Migraine	1	1	1	84
	Osteoarthritis	1	1	1	84
	Paralysis (brain hemorrhage)	1	1	1	84
	Peptic ulcer disease	1	1	1	84
	Pneumonia	1	1	1	84
	Psoriasis	1	1	1	85
	Tuberculosis	1	1	1	84
	Typhoid	1	1	1	84
	Urinary tract infection	1	1	1	84
	Varicose veins	1	1	1	84
	Hepatitis A	1	1	1	84
Metrics	Accuracy		1		3444
	Macro avg		1		3444
	Weighted avg		1		3444

3.2. Stage 2

In the test stage, 1,476 new samples enter the trained model and will be converted into the corresponding disease type. Unlike the training validation, these data were not included in the training process to converge the centroids, so this is an important step to verify whether our KC^{3}I model works practically. As shown in Table 4, the accuracy and F1-score accomplish 100 % and indicate the absolute precision of disease prediction, respected with the reference results.

Figure 5 presents the confusion matrix for inter-subject disease recognition. 41 disease types are represented by the number based on the Dictionary box. All the disease prediction matches the true label that indicates no misclassification between the considered classes. The confusion matrix justifies that the model highly possesses effective performance for disease prediction. The strong point of this approach is that it only requires solely one labelled data per type for training.

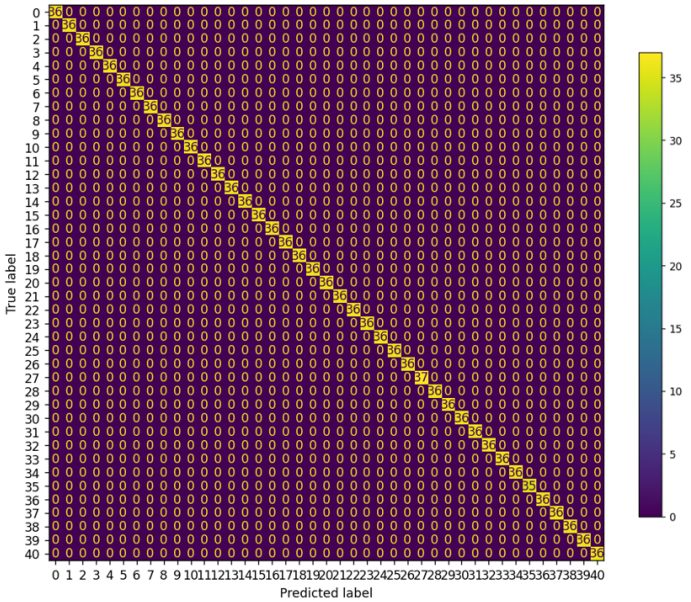

Figure 5. Confusion matrix of test data

Table 4. Classification report of test data

	Disease Type	Precision	Recall	F1-score	Support Number
Disease types	Paroymsal Positional Vertigo	1	1	1	36
	AIDS	1	1	1	36
	Acne	1	1	1	36
	Alcoholic hepatitis	1	1	1	36
	Allergy	1	1	1	36
	Arthritis	1	1	1	36
	Bronchial Asthma	1	1	1	36
	Cervical spondylosis	1	1	1	36
	Chicken pox	1	1	1	36
	Chronic cholestasis	1	1	1	36
	Common Cold	1	1	1	36
	Dengue	1	1	1	36
	Diabetes	1	1	1	36
	Dimorphic hemmorhoids(piles)	1	1	1	36
	Drug Reaction	1	1	1	36
	Fungal infection	1	1	1	36
	GERD	1	1	1	36
	Gastroenteritis	1	1	1	36
	Heart attack	1	1	1	36
	Hepatitis B	1	1	1	36
	Hepatitis C	1	1	1	36
	Hepatitis D	1	1	1	36
	Hepatitis E	1	1	1	36
	Hypertension	1	1	1	36
	Hyperthyroidism	1	1	1	36
	Hypoglycemia	1	1	1	36
	Hypothyroidism	1	1	1	36
	Impetigo	1	1	1	37
	Jaundice	1	1	1	36
	Malaria	1	1	1	36
	Migraine	1	1	1	36
	Osteoarthristis	1	1	1	36
	Paralysis (brain hemorrhage)	1	1	1	36
	Peptic ulcer diseae	1	1	1	36
	Pneumonia	1	1	1	36
	Psoriasis	1	1	1	35
	Tuberculosis	1	1	1	36
	Typhoid	1	1	1	36
	Urinary tract infection	1	1	1	36
	Varicose veins	1	1	1	36
	Hepatitis A	1	1	1	36
Metrics	Accuracy			1	1476
	Macro avg	1	1	1	1476
	Weighted avg	1	1	1	1476

4. FEATURE REDUCTION

Once the model works well with 132 features, the most important symptoms will be assessed to find out the possibility of reducing the number of input features, but still maintaining the absolute efficiency of the system. In this state, all the sample included test and trained data are fed into the model. The 'random testing feature' will show the sufficient number of required features to maintain absolute accuracy.

All the most necessary symptoms can be tracked using the 'Feature importance 'method, an inbuilt class that comes with Tree-Based Classifiers [29]. The extra-trees classifier fits a number of randomized decision trees to the data, as an ensemble learning based on Gini impurity [28]. Particularly, random splits prevent the model from overfitting the data. In this case, there are 100 estimators, with minimum splitting sample of 2 and minimum leaf sample of 1.

To detect the most predominant features of root node, the algorithm calculates how poorly each feature divided the data into the correct class. This calculation measures the impurity of the split, then the feature with the lowest impurity is the most suitable feature for splitting the current node. This process would continue for each subsequent node using the remaining features.

Consider dataset K which contains samples from c classes. The probability of samples belonging to class b at a given node can be denoted as p_b. Then the Gini Impurity is defined as,

$$\text{Gini (K)} = 1 - \sum_{b=1}^{c} p_b^2 \quad (9)$$

The node with uniform class distribution has the highest impurity. The minimum impurity is obtained when all records belong to the same class. Figure 6 shows the 20 most important features as symptoms are demonstrated by using Extra Trees Classifier, which give higher score for the more essential features.

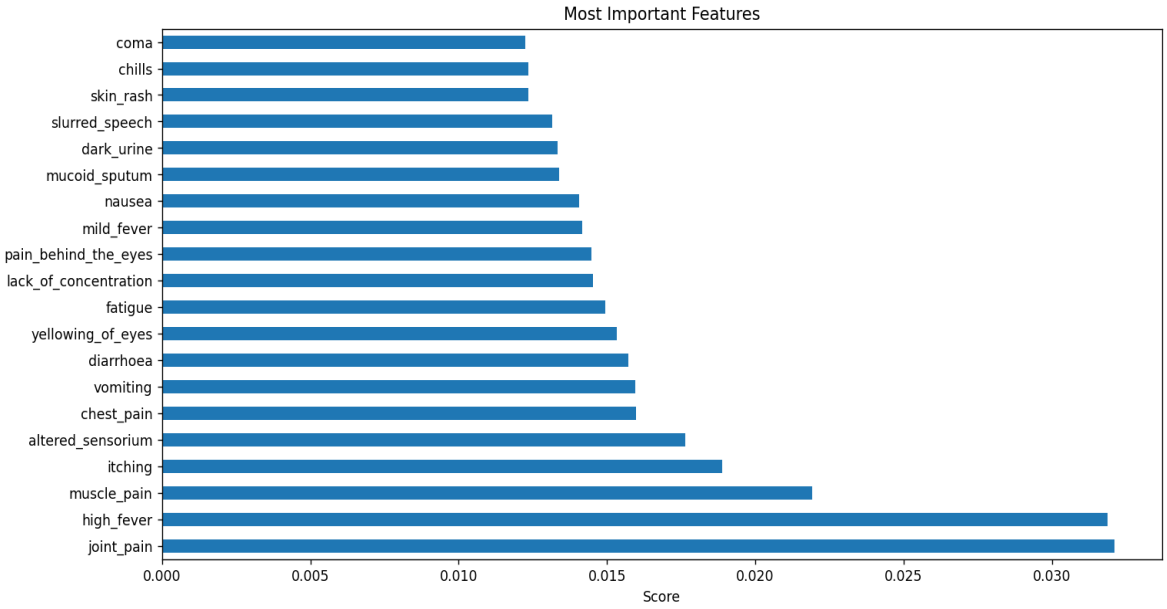

Figure 6. Twenty most important symptoms for disease prediction

Consider M is the number of utilized features for disease prediction. At 1st iteration, M-1 most important features are inserted into the KC^{3}I system. Then, all the predictions are made from the total dataset input to get the accuracy percentage. Next iteration contains the M-2 most important feature. The cycle will repeat until the accuracy is less than 100% at iteration i. At this point, the selected number of features is M -i -1, the feature number of the previous iteration before the operation ends. In this case, the loop is stopped at 115, so 116 most important features can be used for disease prediction.

As a reported in Table 5, about 116 features are necessary to guarantee disease prediction accuracy. Hence, it is possible to remove 16 among 132 symptoms and still guarantee the prediction quality with the KC^{3}I system. If it is forced to cut down further features, the accuracy will continue to drop, as shown in Table 5 and Figure 7.

Table 5. Feature number and equivalent accuracy.

Feature Number	Accuracy
131	100
130	100
129	100
116	100
115	99

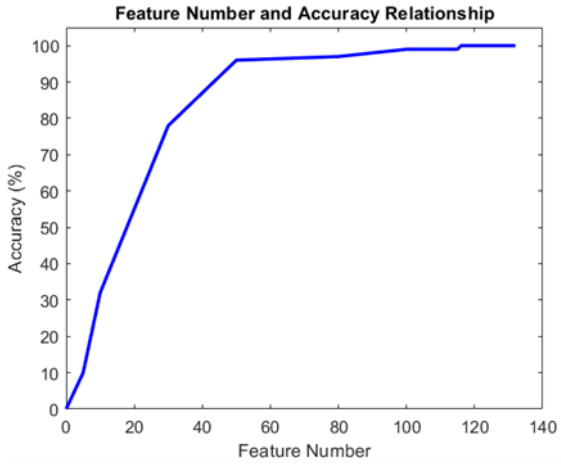

Figure 7. Feature number and accuracy graph

5. CONCLUSION

The main goal of this article is to deeply research and apply the seeded K-means clustering algorithm to the disease prediction. The difference with the applied system KC^{3}I and others is to require only one labelled data per type as the initial centroid for each cluster, and then converges other close-distance samples to its group for training. With the auto-indexed process and dictionary box, the system permanently recognizes the corresponding symptom and each centroid, even after updating the procedure. This technique opens a new way to approach the cases of limited data where the true label is rare and difficult to obtain. This system is close to the idea of unsupervised learning for specific output classification since it just needs solely one known label per type, which is usually possible to collect. Therefore, KC^{3}I is extremely useful to save time in labelling data and is very practical with high accuracy in disease prediction. In addition, the research also shows the possibility of reducing input features but maintaining the absolute efficiency and accuracy of the proposed technique. In the future, we would like to apply the KC^{3}I technique to more circumstances where data have limited input features to observe further pros and cons of the system. From that, the system can be advanced for broad applications in ML field.

ACKNOWLEDGEMENTS

This research was granted by University of Parma through the action "BANDO YIRG UNIPR" co-funded by MUR-Italian Ministry of Universities and Research-D.M. 737/2021-PNR-PNRR-NextGenerationEU to M.M.

REFERENCES

[1] M. Ullrich, A. Küderle, L. Reggi, A. Cereatti, B. M. Eskofier and F. Kluge, "Machine learning-based distinction of left and right foot contacts in lower back inertial sensor gait data," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 5958-5961, doi: 10.1109/EMBC46164.2021.9630653.

[2] E. K. Lee, *Machine learning framework for classification in medicine and biology*, in Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 1–7, doi: 10.1007/978-3-642-01929-6_1.

[3] J. H. Oh, R. Al-Lozi and I. E. Naqa, "Application of Machine Learning Techniques for Prediction of Radiation Pneumonitis in Lung Cancer Patients," 2009 International Conference on Machine Learning and Applications, Miami, FL, USA, 2009, pp. 478-483, doi: 10.1109/ICMLA.2009.118.

[4] G. Jia, H. -K. Lam, S. Ma, Z. Yang, Y. Xu and B. Xiao, "Classification of Electromyographic Hand Gesture Signals Using Modified Fuzzy C-Means Clustering and Two-Step Machine Learning Approach," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 6, pp. 1428-1435, June 2020, doi: 10.1109/TNSRE.2020.2986884.

[5] F. Akhbardeh, F. Vasefi, N. MacKinnon, M. Amini, A. Akhbardeh, and K. Tavakolian, "Classification and assessment of hand arthritis stage using support vector machine," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, 2019, pp. 4080-4083, doi: 10.1109/EMBC.2019.8857022.
[6] P. Singh, S. P. Singh and D. S. Singh, "An introduction and review on machine learning applications in medicine and healthcare," *2019 IEEE Conference on Information and Communication Technology*, Allahabad, India, 2019, pp. 1-6, doi: 10.1109/CICT48419.2019.9066250.
[7] A. Brunhuemer, L. Larcher, P. Seidl, S. Desmettre, J. Kofler, and G. Larcher, "Supervised machine learning classification for short straddles on the S&P500," *Risks*, vol. 10, no. 12, p. 235, 2022., doi: 10.3390/risks10120235.
[8] C. Satinet and F. Fouss, "A supervised machine learning classification framework for clothing products' sustainability," *Sustainability*, vol. 14, no. 3, p. 1334, Jan. 2022, doi: 10.3390/su14031334.
[9] R. Pinky, S. J. Singh and C. Pankaj, "Human Activities Recognition and Monitoring System Using Machine Learning Techniques," *2022 Trends in Electrical, Electronics, Computer Engineering Conference (TEECCON)*, Bengaluru, India, 2022, pp. 62-66, doi: 10.1109/TEECCON54414.2022.9854829.
[10] O. M. Prabowo, K. Mutijarsa and S. H. Supangkat, "Missing data handling using machine learning for human activity recognition on mobile device," *2016 International Conference on ICT For Smart Society (ICISS)*, Surabaya, Indonesia, 2016, pp. 59-62, doi: 10.1109/ICTSS.2016.7792849.
[11] Nisha and P. J. Kaur, "A survey of clustering techniques and algorithms," *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, 2015, pp. 304-307.
[12] R. Krishnamoorthy and S. S. Kumar, "Optimized cluster validation technique for unsupervised clustering techniques," *International Conference on Information Communication and Embedded Systems (ICICES2014)*, Chennai, India, 2014, pp. 1-6, doi: 10.1109/ICICES.2014.7033782.
[13] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Englewood Cliffs, NJ, U.S.A.: Prentice-Hall, 1988.
[14] K. P. Sinaga and M. -S. Yang, "Unsupervised k-means clustering algorithm," in *IEEE Access*, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
[15] M. Alhawarat and M. Hegazi, "Revisiting k-means and topic modeling, a comparison study to cluster Arabic documents," in *IEEE Access*, vol. 6, pp. 42740-42749, 2018, doi: 10.1109/ACCESS.2018.2852648.
[16] Y. Meng, J. Liang, F. Cao, and Y. He, "A new distance with derivative information for functional k-means clustering algorithm," *Information Sciences*, vol. 463–464, pp. 166–185, 2018, doi: 10.1016/j.ins.2018.06.035.
[17] Z. Lv, T. Liu, C. Shi, J. A. Benediktsson and H. Du, "Novel land cover change detection method based on k-means clustering and adaptive majority voting using bitemporal remote sensing images," in *IEEE Access*, vol. 7, pp. 34425-34437, 2019, doi: 10.1109/ACCESS.2019.2892648.
[18] J. Zhu, Z. Jiang, G. D. Evangelidis, C. Zhang, S. Pang, and Z. Li, "Efficient registration of multi-view point sets by K-means clustering," *Information Sciences*, vol. 488, pp. 205–218, 2019, doi: 10.1016/j.ins.2019.03.024.
[19] K. R. Dalal, "Analysing the role of supervised and unsupervised machine learning in IoT," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2020, pp. 75-79, doi: 10.1109/ICESC48915.2020.9155761.
[20] K. K. Jha, R. Jha, A. K. Jha, M. A. M. Hassan, S. K. Yadav and T. Mahesh, "A brief comparison on machine learning algorithms based on various applications: a comprehensive survey," *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Bangalore, India, 2021, pp. 1-5, doi: 10.1109/CSITSS54238.2021.9683524.
[21] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1-130, 2009, doi: 10.2200/S00196ED1V01Y200906AIM006.
[22] P. K. Mallapragada, R. Jin, A. K. Jain and Y. Liu, "SemiBoost: boosting for semi-supervised learning," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2000-2014, Nov. 2009, doi: 10.1109/TPAMI.2008.235.
[23] O. Chapelle, B. Scholkopf and A. Zien, Eds., "Semi-supervised learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]," *in IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542-542, March 2009, doi: 10.1109/TNN.2009.2015974.
[24] E. Bair, *Semi-supervised clustering methods*, Wiley Interdiscip. Rev. Comput. Stat., vol. 5, no. 5, pp. 349–361, 2013, doi: 10.1002/wics.1270.
[25] S. Basu, A. Banerjee, R. Mooney, "Semi-supervised clustering by seeding,*" In Proceedings of 19th International Conference on Machine Learning (ICML-2002*), pp. 19-26 2002.
[26] L. Gu and X. Lu, "Semi-supervised subtractive clustering by seeding," *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, Chongqing, China, 2012, pp. 738-741, doi: 10.1109/FSKD.2012.6234240.
[27] Kaggle, "Disease Prediction Using Machine Learning," Kaggle, Accessed: December 20, 2022. [Online] Available: https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning?resource=download&select=Training.csv.
[28] P. Geurts, D. Ernst., and L. Wehenkel, *Extremely randomized trees*, Machine Learning, vol. 63, no. 1, pp. 3-42, 2006.
[29] *sklearn.ensemble.ExtraTreesClassifier*. Scikit learn. Accesed: January 16, 2023. [Online] Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html.
[30] A. Cheruvu, V. Radhakrishna and N. Rajasekhar, "Using normal distribution to retrieve temporal associations by Euclidean distance," *2017 International Conference on Engineering & MIS (ICEMIS)*, Monastir, Tunisia, 2017, pp. 1-3, doi: 10.1109/ICEMIS.2017.8273101.
[31] "Python," Python.org, Accessed: May 20, 2023. [Online]. Available: https://www.python.org/.
[32] "Scikit-learn," Scikit-learn.org, Accessed: May 25, 2023. [Online]. Available: https://scikit-learn.org/stable/.
[33] M. L. Hoang, A. A. Nkembi, and P. L. Pham, "Real-time risk assessment detection for weak people by parallel training logical execution of a supervised learning system based on an iot wearable MEMS accelerometer," *Sensors*, vol. 23, no. 3, p. 1516, Jan. 2023, doi: 10.3390/s23031516.
[34] M. L. Hoang and A. Pietrosanto, "New artificial intelligence approach to inclination measurement based on MEMS accelerometer," *in IEEE Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 67-77, Feb. 2022, doi: 10.1109/TAI.2021.3105494.
[35] M. Long Hoang and A. Pietrosanto, "Yaw/Heading optimization by machine learning model based on MEMS magnetometer under harsh conditions," Measurement, vol. 193, no. 111013, April 2022, doi: 10.1016/j.measurement.2022.111013.
[36] M. L. Hoang, M. Carratù, V. Paciello and A. Pietrosanto, "A new orientation method for inclinometer based on MEMS accelerometer used in Industry 4.0," *2020 IEEE 18th International Conference on Industrial Informatics (INDIN)*, Warwick,

United Kingdom, 2020, pp. 177-181, doi: 10.1109/INDIN45582.2020.9442189.
[37] M. L. Hoang and A. Pietrosanto, "An effective method on vibration immunity for inclinometer based on MEMS accelerometer," *2020 International Semiconductor Conference (CAS)*, Sinaia, Romania, 2020, pp. 105-108, doi: 10.1109/CAS50358.2020.9267997.

BIOGRAPHIES OF AUTHORS

Dr. Minh Long Hoang is a research fellow at the University of Parma in Dept. of Engineering and Architecture He achieved his Ph.D. in the industrial engineering about "Industry 4.0 oriented enhancement of Inertial Platform performance". He starts working on multiple projects of Machine Learning as a postdoc. His research interests include inertial measurement unit (IMU) sensors, microelectromechanical systems (MEMS), real-time measurements, embedded systems, signal processing, machine learning and deep learning. He can be contacted at email: minhlong.hoang@unipr.it

Prof. Nicola Delmonte is a professor at the University of Parma in Dept. of Engineering and Architecture. He participates at different national and international research projects. Among these, two PRIN (national (Italy) research projects) on power converters development and reliability in harsh environment, and the APOLLO and project in association with the INFN (Istituto Nazionale di Fisica Nucleare, Pavia Section), to simulate and characterize the Main Converter for an upgrade of the Large Hadron Collider (LHC) experiments at CERN. His research interests include design and simulation of power modules and devices, microwave characterization of dielectric thin film and devices, reliability of electronic devices, thermal management of power electronics, PV cells microtechnology, wave energy converters, and smart grids. He has also participated in multiple machine learning projects. He can be contacted at email: nicola.delmonte@unipr.it