Journal of Theoretical and Applied Information Technology
31st December 2021. Vol.99. No 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DNA SEQUENCE ANALYSIS FOR DISEASE PREDICTION AND TREATMENT BASED ON MACHINE LEARNING

[1]Romany M. Farag, [2]M. A. El-Dosuky, [3]M.Z. Rashad

[1]Business informatics Dept. Faculty of Computer and Information, Mansoura University, Egypt

[2,3] Computer Science Dept. Faculty of Computer and Information, Mansoura University, Egypt

E-mail: [1]roumanymesshia@hotmail.com, [2]meldosuky@gmail.com, [3]magdi_z2011@yahoo.com

ABSTRACT

Biomedical data management is crucial in biomedical systems. DNA sequence analysis performed to predict diseases. This paper reviews the most recent machine learning methods as employed in the medical field such as K-Nearest-Neighbor (KNN), Gaussian Process (GP) Classifier, Decision Tree (DT) Classifier, Random Forest (RF) Classifier, Multi-Layer Perceptron (MLP) Classifier, Ada Boost Classifier, Support Vector Machine (SVM) and Deep Learning (DL). The paper applies those methods on a standard DNA dataset. The paper proposes a biomedical data management framework. Then the paper introduces Bag-of-Words (BoW) Random Forest (RF), which achieved 100% accuracy compared to other machine learning methods. Furthermore, this paper introduces a treatment protocol recommendation based on DNA alignment and position matrix.

Keywords: *Disease Prediction, Disease Treatment, DNA Sequence, Machine Learning, Biomedical Data Management*

1. INTRODUCTION

Artificial intelligence techniques are among the best tools used in the health field [1]. Biomedical data management is critical in biomedical systems. DNA sequencing analysis can be performed in order to predict diseases. As shown in Figure 1, there are three main sciences that can be combined to formulate and assist biomedical informatics. The first is machine learning (ML) with which it can classify and predict properties applied by DNA sequencing. Then bioinformatics that combines biological sciences and computer sciences to learn and analyze biological information. Medical informatics helps medicine predict diseases that can affect humans and others [2,3].

First, this paper reviews machine learning methods as employed in the medical field such as K-Nearest-Neighbor (KNN), Gaussian Process Classifier, Decision Tree Classifier, Random Forest Classifier, Multi-Layer Perceptron Classifier (MLP), Ada Boost Classifier, Support Vector Machine, and deep learning. The paper applies those methods on a standard DNA dataset. The paper proposes a biomedical data management framework. Then the paper introduces bag-of-words random forest, which gives 100% accuracy compared to other machine learning methods. The paper also introduces a treatment protocol recommendation based on DNA alignment and position matrix.

In the past period, it has been found that some diseases that occur due to genetic mutations or DNA deviation in some cells have emerged. Therefore, these diseases are fatal and there is no cure for them except with surgery first, and then with chemotherapy. During the short period, some new types of treatment appeared, such as gene therapy. On the other hand, clinical treatment, but there is a common factor between these treatments, which is the disease itself, which is everything types of cancer, but there are some for which surgery is not used [4-6].

We will use artificial intelligence techniques and methods to predict the likelihood of developing cancer. If the possibility of infection is high, the system is activated and everything necessary is prepared through the databases located within medical institutions, hospitals or private treatment centers so that appropriate and faster protocols are made and prepared to avoid the problem of infection or infection of the appropriate treatment protocol through DNA where the work is done comparisons of nucleic acids together until similarity ratios are found between the patient and their ancestors of the same condition, so that the

necessary actions are taken with the highest success rates and without any side effects. As shown in Figure 2, the procedure of using AI and ML to deal with the new applied disease data based medical informatics and the DNA sequence analysis to make a comparisons and analysis of the data based on AI techniques [7]. The major contribution of this paper is listed as follows: -

- The paper proposed a biomedical data management framework to identify specific diseases for DNA sequence analysis.
- A new applied hybrid algorithm based on BoW and RF is utilized to improve the achieved results compared with different ML applied classifiers.
- A treatment protocol is introduces based on DNA alignment and position matrix.

The remaining of the paper is the related work in Section (2), the proposed methodologies is introduced in Section (3), the results and discussions are presented in Section (4), finally, the conclusion and future work are in Section (5).

2. RELATED WORK

In this section, we introduced the most common ML approaches utilized in medical applications such as KNN, GP, DT, RF, MLP, Ada boost, SVM, and DL for classification and regression of medical applications. Furthermore, we attempt to overview the DNA Sequence Alignment and Analysis including DNA sequence alignment and prediction.

2.1 Machine Learning in Medical Field

2.1.1 K-Nearest-Neighbor (KNN)

K-Nearest-Neighbor (KNN) has been used in solving many problems of early diagnosis of diseases [1]. Different classifications are combined to be more efficient than the individual classifications. KNN was used in determining the difference between the position of falling and lying down for a person and determining the critical time difference to discover falls accidents [8]. It is also used to recognize the enlargement of the thyroid gland and the early detection of the disease [9]. The discovery of mental diseases and the treatment are dealt with through the use of existing algorithms such as KNN and the use of electrical brain sensors [10]. Data mining tools are used inside medical databases for analyzing them [11]. The classification is made closest to the similar classes with which the method is used.

2.1.2 Gaussian Process Classifier

One of the best tools used to make a classification for a group of students who suffer from stress is the Gaussian classification [12]. The experiment was done to reduce stress to relieve stress. The experiment succeeded and the accuracy rate was 94%. The Gaussian tool was used in the early and inexpensive detection without any surgeries for a disease, and the experiment was very effective, as the tool was used on 65 people, including 15 healthy people and 50 patients, and the results were very impressive. It was recommended to use this tool for the initial examination and disease prediction [13]. A classification of the EEGs of the human brain was done through the use of the Gaussian classifier [14]. Cases were classified by monitoring the signals within the brain. It was very useful and effective in making analysis. Vision has been developed in medical robots that perform surgeries by using Gaussian classification by analyzing three-dimensional images of the organism [15]. Gaussian processes have also been used to improve the parameters of predictive performance based on design algorithms [16].

2.1.3 Decision Tree Classifier

Among the distinct techniques used in making the classifications is the decision tree technique, where the J4.8 algorithm was designed on the basis of gain ratios and binary estimates [17]. The performance of these algorithms was very superior in diagnosing heart disease. The decision tree and neural networks were used in digging into the data for research [18]. On the other hand, the decision tree was used to search in the data to find out the common reason among women for choosing a method of contraception [19]. The decision tree was also very effective in the predictive value of dengue and white blood fever in the adult clinical environment [20]. The overall accuracy rate of the decision tree was good as it was sensitive to specific tests. Decision tree is one of the most basic pillars in early disease as hybrid systems are designed to deal with breast cancer and best way to treat it [21].

2.1.4 Random Forest Classifier

This method is used to predict the disease based on the patient's history, where samples are classified from patients into eight disease categories [22]. This method was very effective in addition to the RF method that has the advantage of calculating all the variables separately within the classification

process. By using the Internet of Things, healthcare is increased by developing health systems analyzes using IOT and by using improved algorithms to obtain a better classification than other methods, and also by using RF, the accuracy was high at 94.2% [23]. And by means of the Internet of things, different techniques are applied, from machine learning to the health data contained within the cloud databases [24]. Through the application of different technologies, a correct decision is made in a very large proportion through the prediction of disease. Through the use of monitoring devices and remote sensors that are worn, diseases are classified. For different diseases, the usual and boldest methods were used in making the required classifications. After that the focus was on analyzing the dimensions of the features and performing the comparative analysis such as error measures, ROC curves and confusion matrix. Diseases were accurately detected and predicted with very excellent efficacy [25]. RF techniques have been used in performing EEG analyzes to identify and predict epileptic seizures using standard EEG and they have been very effective [26].

2.1.5 Multi-Layer Perceptron Classifier

EEG technology was used and analyzes were made to predict and identify Alzheimer's disease for the elderly, and then classifications are made for these patients, with votes for the majority and decision templates [27]. Through the presence of a database within the health system, which are multi-layered algorithms in order to discover the causes of diabetes, where its causes were hereditary or due to medical negligence and other causes. The disease was also classified, and a model was superior to all other classification models and had excellent accuracy [28]. Artificial intelligence techniques with the Internet of Things (IoT) have been used in the early diagnosis and remote follow-up of patients in many fatal diseases, and the experiments in diagnosis and prediction of the disease were very effective and excellent as well, and this newly developed health model has helped doctors very effectively [29]. MLP was used as classifiers in machine learning to classify groups in the voting technique to evaluate human activity by wireless sensors and the techniques used showed remarkable superiority over the individual classifiers [30]. In machine learning, attacks on training data can occur, leading to fatal errors in appropriate decision-making in the healthcare field. Six machine learning algorithms and five health care data sets are identified where countermeasures are provided against general attacks and detection [31].

2.1.6 Ada Boost Classifier

The use of artificial intelligence techniques, especially machine learning techniques, has significantly reduced the burden on health systems, as they are burdened all over the world with successive waves of the Corona virus, and with the increasing number of infected people, machine learning techniques were very effective through proactive treatment protocols [32]. The performance of most of the evaluated data in particular for more complex classification tasks was improved through the use of the Ada Boost [33]. A decision tree was used on a set of classifiers and its control on diabetic patients achieved the highest standards in performance compared to other algorithms [34]. Data science is based on early prediction of breast cancer and the use of data mining techniques [35]. Ada Boost helped improve the classification process of a dangerous disease that causes death [36].

2.1.7 Support Vector Machine

There are an infinite number of hyperplanes. This method attempts to find a hyper-plane that maximizes the gap between support vectors which are data points on the boundaries. It transforms the space into another space that allows for easily separating non-linear problems.

Through the use of smart devices built with sensors inside, a classification and exploration are made within the incoming data, and through machine learning the daily routine is defined through the use of SVM algorithms and algorithms, where if a specific error occurs within the reading, a distinction is made between technical error and actual error and through These procedures are necessary for patient [37].

And through the massive transformation of digital data, artificial intelligence techniques, especially SVM, and its algorithms were used to develop the treatment industry to improve health life and patient service [38].

Also, a development was made through SV to improve the storage and follow-up of patients through cloud storage and follow-up each A case by monitoring its data [39].

And through a study conducted on 249 employees who work within the health sector, so that the researchers could know the effect of the

administration's treatment, salaries and others on their performance within the health sector, and SV algorithms were used to monitor these employees. Excellent effectiveness in obtaining the desired results [40].

To obtain the signal that was previously processed to obtain the feature, SVM was used on the signals from the heart, noise was removed, and the sound of beats was filtered, and the use of SVM Classifier was one of the best tools that were used [41].

2.1.8 Deep learning

Deep learning techniques have been used in the fields of health care through computer vision, where language is processed, the method of exploring the design of medical systems and the design of robots for accurate surgeries and reviews in genomics [42].

The biggest challenge in this era was the presence of a huge amount of unused vital data, and deep learning techniques were applied to provide a comprehensive and insightful view in order to form a comprehensive perspective of bioinformatics [43]. Deep learning has been used to achieve great success in solving problems facing researchers in medical informatics, as neural networks and their graphics and automatic coding have been used and the latest technologies have been used in this field [44]. Hence, deep learning has shown an enormous and massive explosion in application to biomedical informatics, the discovery of overlapping and intertwining relationships within the genome, and the clarification of some applications that use deep learning and predicting future research in bioinformatics [45].

Artificial intelligence techniques, specifically deep learning, have been used, algorithmic improvements have been used to predict automatically, high-quality features and semantic interpretation have been provided through input data, and the focus has been on vital and medical data [46]. All texts and images were used to study proteins and cells to see the inherent environment within the data, where through the use of deep learning techniques, a comprehensive survey of many groups of data was reviewed and a comprehensive survey was made for many groups of data, and many related processes and different dimensions worked, which led to the improvement of the results and increased its efficiency as in [47].

2.2 DNA Sequence Alignment and Analysis

2.2.1 DNA Sequence Alignment

DNA alignment processes have been used to identify genetic mutations and mutations. Where the jumps made from the process of DNA modifications, for example, Corona Virus, are tracked with its modifications [48-50]. A multi-layer realtor was used with ML machine learning processes with NW to reach a high accuracy rate of 99.70%, where the Adam optimizer was used, and 2912 GB of cell updates were reached per second. It has been applied to two real DNA sequences with a length of 4.1 million nucleotides. The number of steps used in calculating the algorithm have been balanced to accelerate performance with the use of ML [51]. DETCCTOR (INSIDER) has been developed to identify internal DNA sequence information. Where data analyzes are done so that we can know the sequences that were entered on the original sequence from the alien by knowing the large shifts in the signatures of K.mer to identify the biological threats resulting from genetic modification by inserting modified sequences to a specific sequence and the work was done Experiment with wild yeast, and the experience has proven an impressive performance [52]. A mathematical method has been developed to switch from the traditional methods that produce statistics to the innovative method until the regions that are located at the highest sequences are discovered so that cloning from these sequences can be made [53]. The problem of discovering places of DNA, meaning nucleic acids that lack oxygen, was solved using neural networks and k-MER, and through the training received by the machine to discover places of cloning and improved proteins, and the performance of the experiment was 99.3% [54]. Some storage problems for DNA were solved by providing some DNA pop-up cards, the data was written in the form of pre-determined cracks on the site of the original double-stranded DNA backbone, and then this encoded data is reconstructed through high-throughput sequencing [55]. In addition, high-quality sequences were used on ancient DNA molecules, where comparisons are made between the good sequences and the old sequences to identify the ancient communities and establish specialized studies on the samples. Bowtie2 was used to evaluate, and recommended maps were drawn until the sample was preserved [56].

2.2.2 DNA Sequence Analysis for Disease Prediction

Studies have been done on genetic and genetic diseases and they found that the basis of these diseases are humans, and through clinical trials and tracking these diseases through genetic medicine and statistics, the full potential of these experiments was realized [57]. Through data analysis and modeling, diseases that may infect plants are predicted, and then poor areas are affected, and thus famines occur, and here comes the importance of predicting diseases and epidemics, preparing treatments, and then solving many problems [58]. Through deep learning, a new component was used, the informer, which is capable of integrating long-term interactions up to 100 kb. Genome [59]. Clinical trials for coronary artery disease have been done and it was found that age is one of the most important factors, but the genetic factor is one of the most important causes of this disease, and then the classification process was done based on genetic and genetic risks, and the genetic risks were stratified with very good goals and features. Its presence is innate in an individual's DNA. In addition, through inexpensive stratification that can be done worldwide for early prevention of this disease [60].

3. PROPOSED METHODOLOGY

In this section we present biomedical data management framework as shown in Figure 3 by which the data from several databases are collected via Extract, Transform and Load (ETL) process into data warehouse. Then, slicing the data warehouse into targeted data. Then selecting rows and projecting columns yields scoped data, which are fed into prediction (machine learning methods).

As investigated in Figure 4 the proposed flowchart which is based on position matrix [48]. The matrix is to be used to find similarity between the DNA until it is modified to have accurate results. The bonds between C-G are 3 hydrogen bonds. This makes these bonds have more stability compared with the bonds between A and T. Gaps are added accordingly to indicate this dominance between C and G.

Moreover, in Figure 5 the bag-of-words based random forest are presented. Bag-of-words are generated at different k value in k-merization. It is observed that bag-of-words based random forest is the most accurate machine learning method.

4. RESULTS AND DISCUSSIONS

The used dataset is called E. coli promoter gene sequences [49]. It has 59 attributes as follows: the class (positive or negative), instance name, and 57 sequential nucleotide positions. It has 106 DNA sequences, with 50% class distribution (53 positive instances, 53 negative instances). Table 1 shows the first 10 instances and the Gene sequences for each class.

It is clear from Table 2 and Figure 6 that comparison between machine learning methods investigated the utilization of BoW with RF achieves superior results reached to 100% in terms of the accuracy. As shown in Table 1 the DT with d=4, d=5, and d=6 produces accuracies 73.75%, 76.07, and 82.30%, respectively. While the utilization of SVM with Radial Basis Function (RBF) produces 95.50% compared with SVM Sigmoid 90.00% and SVM Linear 88.75%. We further compare the results with Neural networks with a learning rates α between 1, 0.8, and 0.9 are 90.00, 90.00, 88.75%, respectively. The KNN produces 91.25%, 82.5%, and 87.5% for K=2, 3, and 4, respectively. The classical RF achieves 74.80% and Naïve Bayes (NB) was 85.00%. In the other hand the Ada Boost produces 88.75% and the GP 85.70%. Finally, the DL based Convolutional Neural Networks (CNN) achieved 94.00%. That is all compared with the BoW + RF that archives 100% accuracy.

A comparison between Patients' DNA and the corresponding Groups for each patient such that we have three Groups A, B, and C are shown in Table 3. While in Figure 7. show the accuracy of matching of DNA. As shown in Table 3, after comparing nucleic acids and finding similarity ratios between patients, it was found that group C had the highest similarity rate between the new patient and the old patients. The similarity percentage as shown was 4.657%. The statistics also show the degrees of accuracy after modifying the algorithm used and it was excellent with a percentage of 95.342%, which shows the degree of efficiency of the system used in all the required tasks, whether it is prediction or finding the appropriate protocol.

5. CONCLUSION AND FUTURE WORK

Most common Machine learning approaches are utilized in this work to predict whether the patients are seeking or not in terms of the DNA sequences. The treatment protocol is introduced for the sored Datawarehouse in the DNA sequences by determining the similarity between patients in the

database and groups. A comparison between machine learning methods is conducted on a standard DNA dataset. It is observed that bag-of-words based Random Forest is the most accurate machine learning method. Comparing the nucleic acids and finding similarity ratios between patients are the major challenge of this study. Moreover, it was found that group C had the highest similarity rate between the new patient and the old patients. The Future directions may consider big data analysis or applying a secure environment for DNA sequence analysis and alignment framework.

REFRENCES:

[1] Shouman M, Turner T, Stocker R. Applying k-nearest neighbour in diagnosing heart disease patients. International Journal of Information and Education Technology. 2012 Jun 1;2(3):220-3.

[2] Bernstam, E. V., Smith, J. W., & Johnson, T. R. (2010). What is biomedical informatics?. Journal of biomedical informatics, 43(1), 104-110.

[3] Friedman, C. P., & Wyatt, J. (2005). Evaluation methods in biomedical informatics. Springer Science & Business Media.

[4] Hansen, Wendy LJ, Cathrien A. Bruggeman, and Petra FG Wolffs. "Evaluation of new preanalysis sample treatment tools and DNA isolation protocols to improve bacterial pathogen detection in whole blood." Journal of clinical microbiology 47.8 (2009): 2629-2631.

[5] Chiu, R. W., Poon, L. L., Lau, T. K., Leung, T. N., Wong, E. M., & Lo, Y. D. (2001). Effects of blood-processing protocols on fetal and total DNA quantification in maternal plasma. Clinical chemistry, 47(9), 1607-1613.

[6] Zhang, T., Tian, T., Zhou, R., Li, S., Ma, W., Zhang, Y., ... & Lin, Y. (2020). Design, fabrication and applications of tetrahedral DNA nanostructure-based multifunctional complexes in drug delivery and biomedical treatment. Nature Protocols, 15(8), 2728-2757.

[7] Narayanan, A., Keedwell, E. C., & Olsson, B. (2002). Artificial intelligence techniques for bioinformatics. Applied bioinformatics, 1, 191-222.

[8] Liu CL, Lee CH, Lin PM. A fall detection system using k-nearest neighbor classifier. Expert systems with applications. 2010 Oct 1;37(10):7174-81.

[9] Chandel K, Kunwar V, Sabitha S, Choudhury T, Mukherjee S. A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. CSI transactions on ICT. 2016 Dec 1;4(2-4):313-9.

[10] Chandel K, Kunwar V, Sabitha S, Choudhury T, Mukherjee S. A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. CSI transactions on ICT. 2016 Dec 1;4(2-4):313-9.

[11] Khamis HS, Cheruiyot KW, Kimani S. Application of k-nearest neighbour classification in medical data mining. International Journal of Information and Communication Technology Research. 2014 Apr;4(4).

[12] Desai R, Porob P, Rebelo P, Edla DR, Bablani A. EEG Data Classification for Mental State Analysis Using Wavelet Packet Transform and Gaussian Process Classifier. Wireless Personal Communications. 2020 Aug 6:1-21.

[13] Shashikant R, Chaskar U, Phadke L, Patil C. Gaussian process-based kernel as a diagnostic model for prediction of type 2 diabetes mellitus risk using non-linear heart rate variability features. Biomedical Engineering Letters. 2021 Jun 25:1-4.

[14] Wang B, Wan F, Mak PU, Mak PI, Vai MI. EEG signals classification for brain computer interfaces based on Gaussian process classifier. In2009 7th International Conference on Information, Communications and Signal Processing (ICICS) 2009 Dec 8 (pp. 1-5). IEEE.

[15] Hu J, Sun Y, Li G, Jiang G, Tao B. Probability analysis for grasp planning facing the field of medical robotics. Measurement. 2019 Jul 1;141:227-34.

[16] Sundararajan S, Keerthi SS. Predictive Approaches For Gaussian Process Classifier Model Selection September 10, 2008.

[17] Shouman M, Turner T, Stocker R. Using decision tree for diagnosing heart disease patients. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 2011 Dec 1 (pp. 23-30).

[18] Srinivas K, Rani BK, Govrdhan A. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE). 2010 Feb;2(02):250-5.

[19] Bach MP, Cosic D. Data mining usage in health care management: literature survey and decision tree application. Medicinski glasnik. 2008 Jan 1;5(1):57-64.

[20] Lee VJ, Lye DC, Sun Y, Leo YS. Decision tree algorithm in deciding hospitalization for adult

patients with dengue haemorrhagic fever in Singapore. Tropical Medicine & International Health. 2009 Sep;14(9):1154-9.

[21] Lavanya D, Rani KU. Ensemble decision tree classifier for breast cancer data. International Journal of Information Technology Convergence and Services. 2012 Feb 1;2(1):17.

[22] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making. 2011 Dec 1;11(1):51.

[23] Lakshmanaprabu SK, Shankar K, Ilayaraja M, Nasir AW, Vijayakumar V, Chilamkurti N. Random forest for big data classification in the internet of things using optimal features. International journal of machine learning and cybernetics. 2019 Oct 1;10(10):2609-18.

[24] Kaur P, Kumar R, Kumar M. A healthcare monitoring system using random forest and internet of things (IoT). Multimedia Tools and Applications. 2019 Jul 30;78(14):19905-16.

[25] Salih AS, Abraham A. Intelligent Decision Support for Real Time Health Care Monitoring System. InAfro-European Conference for Industrial Advancement 2015 (pp. 183-192). Springer, Cham.

[26] Mursalin M, Zhang Y, Chen Y, Chawla NV. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. Neurocomputing. 2017 Jun 7;241:204-14.

[27] Stepenosky N, Green D, Kounios J, Clark CM, Polikar R. Majority vote and decision template based ensemble classifiers trained on event related potentials for early diagnosis of Alzheimer's disease. In2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings 2006 May 14 (Vol. 5, pp. V-V). IEEE

[28] Mishra S, Tripathy HK, Mallick PK, Bhoi AK, Barsocchi P. EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis. Sensors. 2020 Jan;20(14):4036.

[29] Kishor A, Chakraborty C. Artificial intelligence and internet of things based healthcare 4.0 monitoring system. Wireless Personal Communications. 2021 Jul 3:1-7.

[30] Mozaffari-Kermani M, Sur-Kolay S, Raghunathan A, Jha NK. Systematic poisoning attacks on and defenses for machine learning in healthcare. IEEE journal of biomedical and health informatics. 2014 Jul 30;19(6):1893-905.

[31] Mohd Azmi MS, Sulaiman MN. Accelerator-based human activity recognition using voting technique with NBTree and MLP classifiers. International Journal on Advanced Science, Engineering and Information Technology. 2017 Jan 1;7(1):146-52.

[32] Khan K, Ramsahai E. Maintaining proper health records improves machine learning predictions for novel 2019-nCoV. BMC Medical Informatics and Decision Making. 2021 Dec;21(1):1-3

[33] Reiss A, Hendeby G, Stricker D. Confidence-based multiclass AdaBoost for physical activity monitoring. InProceedings of the 2013 International Symposium on Wearable Computers 2013 Sep 8 (pp. 13-20).

[34] Kelarev AV, Stranieri A, Yearwood JL, Jelinek HF. Empirical study of decision trees and ensemble classifiers for monitoring of diabetes patients in pervasive healthcare. In2012 15th International Conference on Network-Based Information Systems 2012 Sep 26 (pp. 441-446). IEEE.

[35] Kumar V, Mishra BK, Mazzara M, Thanh DN, Verma A. Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. In Advances in Data Science and Management 2020 (pp. 435-442). Springer, Singapore.

[36] Wibawa MS, Maysanjaya IM, Putra IM. Boosted classifier and features selection for enhancing chronic kidney disease diagnose. In2017 5th International Conference on Cyber and IT Service Management (CITSM) 2017 Aug 8 (pp. 1-6). IEEE.

[37] Salem O, Guerassimov A, Mehaoua A, Marcus A, Furht B. Anomaly detection in medical wireless sensor networks using SVM and linear regression models. International Journal of E-Health and Medical Communications (IJEHMC). 2014 Jan 1;5(1):20-45.

[38] Harimoorthy K, Thangavelu M. Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. Journal of Ambient Intelligence and Humanized Computing. 2020 Jan 2:1-9.

[39] Thanigaivasan V, Narayanan SJ, Iyengar SN, Ch N. Analysis of parallel SVM based classification technique on healthcare using big data management in cloud storage. Recent Patents on Computer Science. 2018 Aug 1;11(3):169-78.

[40] Kuzey C. Impact of Health Care Employees' Job Satisfaction On Organizational Performance

Support Vector Machine Approach. European Journal of Economic & Political Studies. 2012 Jun 1;5(1).

[41] Venkatesan C, Karthigaikumar P, Paul A, Satheeskumaran S, Kumar RJ. ECG signal preprocessing and SVM classifier-based abnormality detection in remote healthcare applications. IEEE Access. 2018 Jan 22;6:9767-73.

[42] Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology. 2013 Oct 31;5(5):241-66.

[43] Chow R, Zhong W, Blackmon M, Stolz R, Dowell M. An efficient SVM-GA feature selection model for large healthcare databases. InProceedings of the 10th annual conference on Genetic and evolutionary computation 2008 Jul 12 (pp. 1373-1380).

[44] Alanazi HO, Abdullah AH, Qureshi KN. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. Journal of medical systems. 2017 Apr 1;41(4):69.

[45] Hossain MS, Muhammad G. Healthcare big data voice pathology assessment framework. iEEE Access. 2016 Nov 8;4:7806-15.

[46] Liu X, Liu L, Simske SJ, Liu J. Human daily activity recognition for healthcare using wearable and visual sensing data. In2016 IEEE International Conference on Healthcare Informatics (ICHI) 2016 Oct 4 (pp. 24-31). IEEE

[47] Karim MR, Beyan O, Zappa A, Costa IG, Rebholz-Schuhmann D, Cochez M, Decker S. Deep learning-based clustering approaches for bioinformatics. Briefings in Bioinformatics. 2020 Feb 2.

[48] Shehab S, Shohdy S, Keshk AE. Pomsa: An efficient and precise position-based multiple sequence alignment technique. arXiv preprint arXiv:1708.01508. 2017 Aug 3.

[49] https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Promoter+Gene+Sequences) Accessed January 14, 2021

[50] Park K, Lim S. A multilayer secure biomedical data management system for remotely managing a very large number of diverse personal healthcare devices. BioMed research international. 2015 Jan 1;2015.

[51] Tay AP, Hosking B, Hosking C, Bauer DC, Wilson LO. INSIDER: alignment-free detection of foreign DNA sequences. Computational and structural biotechnology journal. 2021 Jan 1;19:3810-6.

[52] Rashed AE, Amer HM, El-Seddek M, Moustafa HE. Sequence Alignment Using Machine Learning-Based Needleman–Wunsch Algorithm. IEEE Access. 2021 Jul 26;9:109522-35.

[53] Korotkov EV, Suvorova YM, Kostenko DO, Korotkova MA. Multiple Alignment of Promoter Sequences from the Arabidopsis thaliana L. Genome. Genes. 2021 Feb;12(2):135.

[54] Morgenstern B. Sequence comparison without alignment: The SpaM approaches. InMultiple Sequence Alignment 2021 (pp. 121-134). Humana, New York, NY.

[55] Tabatabaei SK, Wang B, Athreya NB, Enghiad B, Hernandez AG, Fields CJ, Leburton JP, Soloveichik D, Zhao H, Milenkovic O. DNA punch cards for storing data on native DNA sequences via enzymatic nicking. Nature communications. 2020 Apr 8;11(1):1-0

[56] Poullet M, Orlando L. Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes. Frontiers in Ecology and Evolution. 2020 May 6;8:105.

[57] Benton ML, Abraham A, LaBella AL, Abbot P, Rokas A, Capra JA. The influence of evolutionary history on human health and disease. Nature Reviews Genetics. 2021 May;22(5):269-83.

[58] Ristaino JB, Anderson PK, Bebber DP, Brauman KA, Cunniffe NJ, Fedoroff NV, Finegold C, Garrett KA, Gilligan CA, Jones CM, Martin MD. The persistent threat of emerging plant disease pandemics to global food security. Proceedings of the National Academy of Sciences. 2021 Jun 8;118(23).

[59] Avsec Z, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. bioRxiv. 2021 Jan 1.

[60] Roberts R, Chang CC, Hadley T. Genetic risk stratification: a paradigm shift in prevention of coronary artery disease. JACC: Basic to Translational Science. 2021 Mar 1;6(3):287-304

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

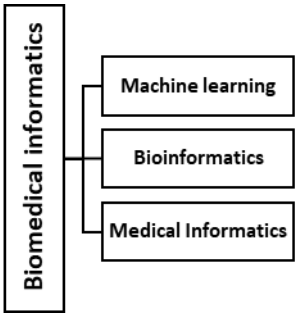

Figure 1: The taxonomy of Biomedical Informatics

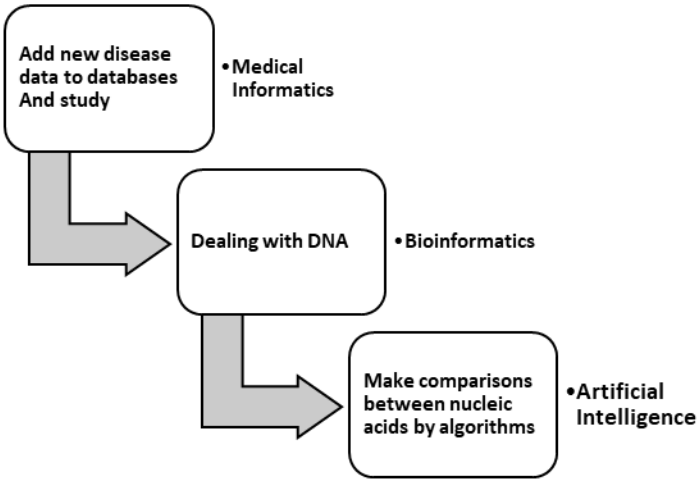

Figure 2. The Bioinformatics steps to make a comparison between nucleic and acids using AI.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

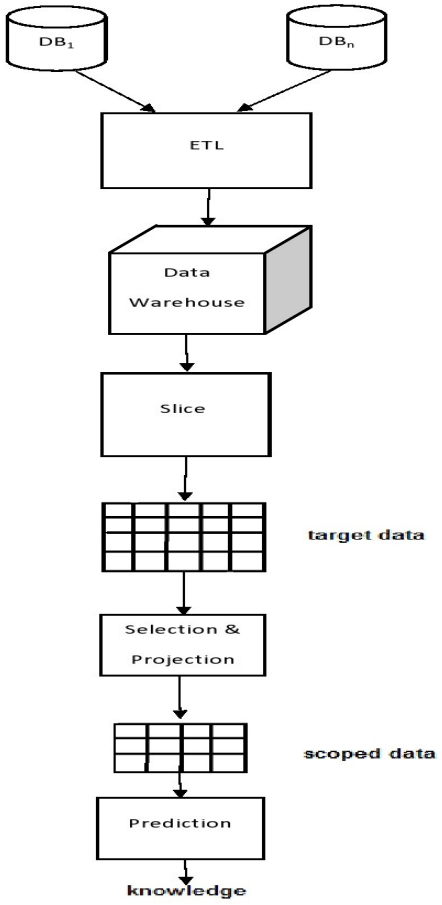

Figure 3. The proposed biomedical data management framework

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

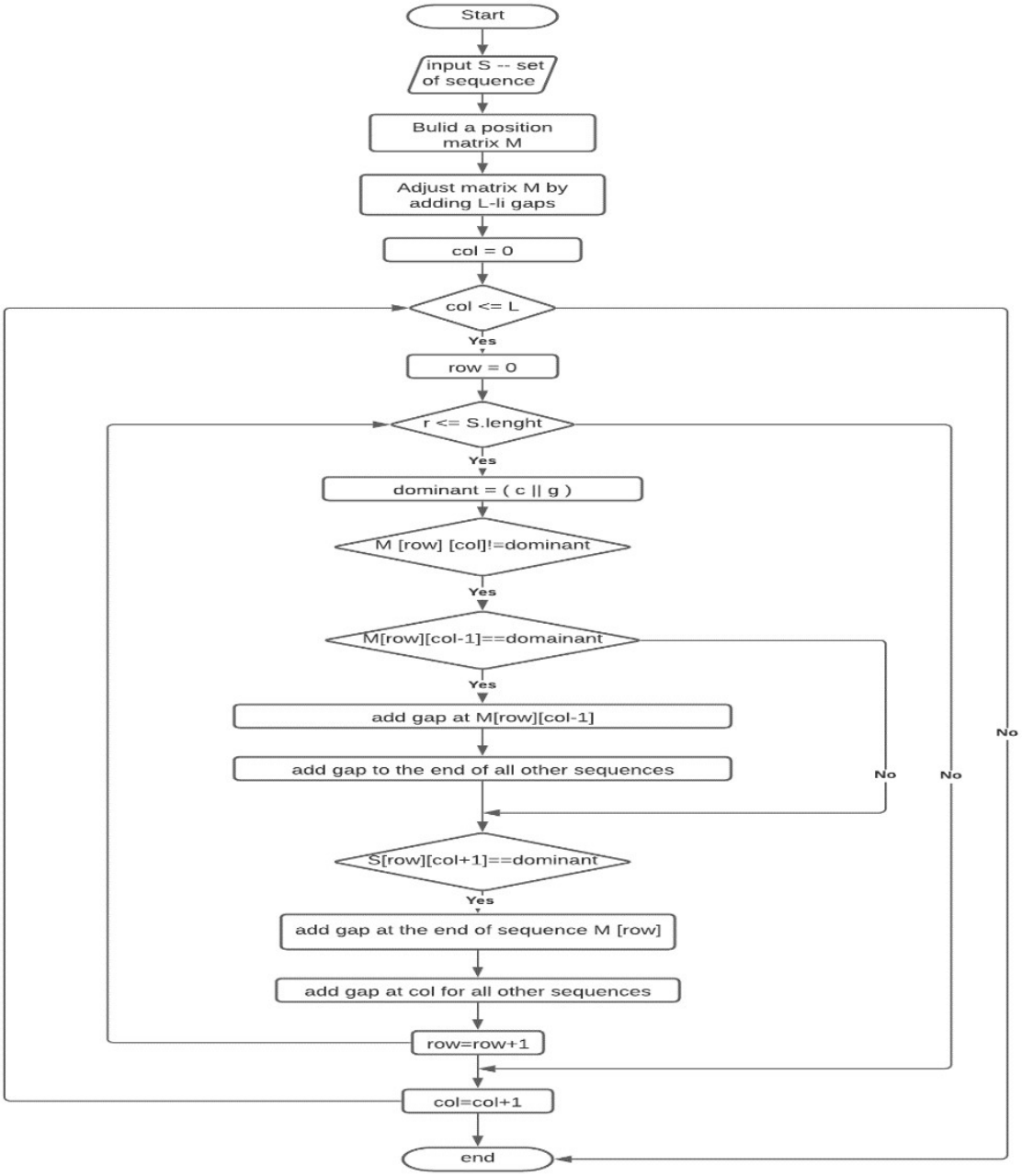

Figure 4. The general flow chart of the proposed method

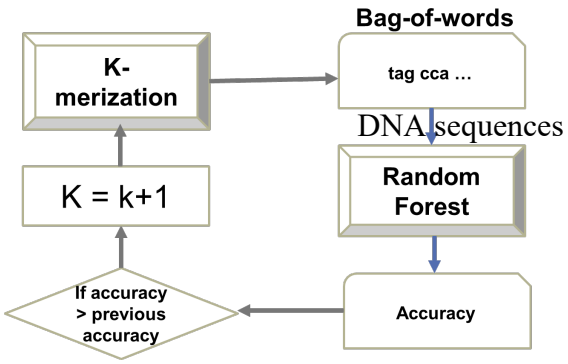

Figure 5 The proposed Bag-of-Words Based Random Forest

Table 1. The most common ten instances and the corresponding Gene sequences

Class	Instance name	Gene sequence
+	S10	tactagcaatacgcttgcgttcggtggttaagtatgtataatgcgc...
+	AMPC	tgctatcctgacagttgtcacgctgattggtgtcgttacaatctaa...
+	AROH	gtactagagaactagtgcattagcttatttttttgttatcatgcta...
+	DEOP2	aattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata...
+	LEU1_TRNA	tcgataattaactattgacgaaaagctgaaaaccactagaatgcgc...
+	MALEFG	aggggcaaggaggatggaaagaggttgccgtataaagaaactagag...
+	MALK	caggggggtggaggatttaagccatctcctgatgacgcatagtcagc...
+	RECA	tttctacaaaacacttgatactgtatgagcatacagtataattgct...
+	RPOB	cgacttaatatactgcgacaggacgtccgttctgtgtaaatcgcaa...
+	RRNAB_P1	ttttaaatttcctcttgtcaggccggaataactccctataatgcgc...

Table 2 Comparison Between Machine Learning Methods

Method	Accuracy (%)
Decision Tree, d=4	73.57
Decision Tree, d=5	76.07
Decision Tree, d=6	82.30
SVM Sigmoid	90.00
SVM Linear	88.75
SVM RBF	92.50
Neural Net, α =1	90.00
Neural Net, α=0.8	90.00
Neural Net, α =0.9	88.75
KNN, k=2	91.25
KNN, k=3	82.50
KNN, k=4	87.50
Random Forest	74.80
Naive Bayes	85.00
Ada Boost	88.75
Gaussian Process	85.70
Deep learning (CNN)	94.00
Bag-of-words based Random Forest	100.00

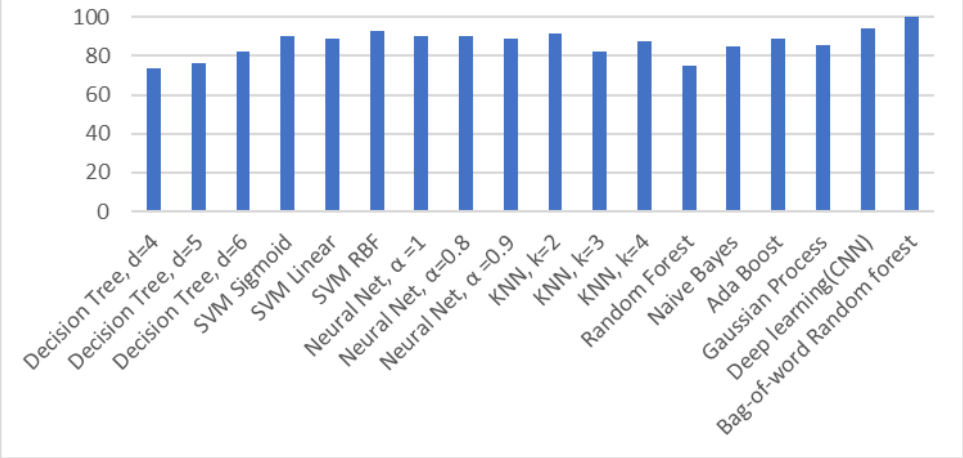

Figure 6. The Comparison between machine learning methods

Table 3. The accuracies of DNA matching between Groups and patients with the similarity and final score values.

Groups	Patient	Similarity	Final Score	Accuracy
A	A1,A2	16	364	95.604%
B	B1,B2	11	319	96.551%
C	C1,C2	32	687	95.342%

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

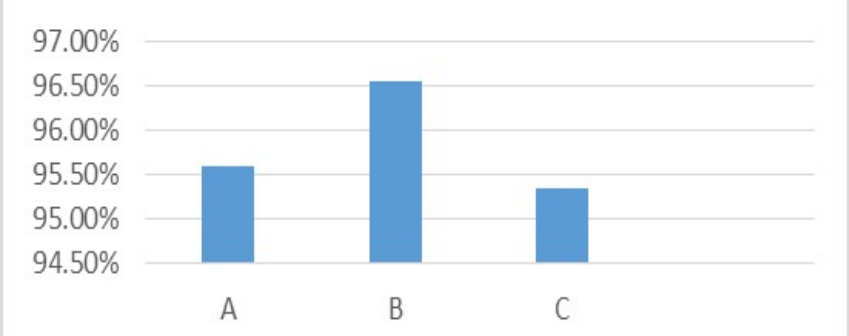

Figure 5 Accuracy matching of DNA