# Mutation Detection in Genes Sequence Using Machine Learning

Syed Muhammad Shehryar
Department of Computer Sciences,
Bahria University Lahore Campus
shehriarhaider98@gmail.com

Muhammad Ammar Shahid
Department of Computer Sciences,
Bahria University Lahore Campus
ammarshahid045@gmail.com

Asghar Ali Shah
Department of Computer Sciences,
Bahria University Lahore Campus
alishahsadiq@gmail.com

2021 International Conference on Innovative Computing (ICIC) | 978-1-6654-0091-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIC53490.2021.9693088

***Abstract*— Cancer has been identified as a serious genetic disorder which cause numerous deaths every year. Late diagnosis of cancer is one of the major cause of deaths. Mutation is the disturbance in gene sequence and mutated genes are crucial for cancer growth. This study develops a machine learning model for the solution of mutation detection problem in genes sequence which may help in diagnosing the cancer at early stages. The aim of this system is to detect the mutation in gene sequences whether it is mutated or not. To detect mutation in gene sequences, several techniques and models of machine learning has been considered such as SVM, Logistic Regression and Linear Discriminant Analysis to achieve better accuracy.**

***Keywords—cancer, mutation, genes sequence***

## I. INTRODUCTION

Diseases are ancient as humanity is on the earth. With the passage of time science and technology put their efforts in the field of medical. Development in technology is making easy to diagnose and cure a disease. Cancer is a crucial disorder in the modern era. Millions of people died every year because of the late diagnose of cancer. While mutation in genes is a cause of cancer. Cancer is the abnormal growth of cells and mutation is the disturbance in gene sequence. There is a need of such method which helps to detect mutation so it might be helpful in diagnosing cancer before it's occurrence or at initial stages. Cancer is recognized as a complex disorder which causes due to abnormal growth of cells. Several cancer patients died every year due to late diagnosis of disease. One of the major causes of cancer is mutation. Mutation is the disturbance in gene sequence. Somatic mutation is one of the types of mutation and it's the common cause of occurrence of cancer [1]. This study proposes a solution to detect mutation in human genes and the mutation detection could be helpful in identifying cancer before its occurrence or at early stages. Diagnosing cancer at initial phases might be useful for proper treatment of cancer and this could be a saviour of many lives.

### *A. Genomes and Genes*

Genomes are also referred to as the library of genes sequences. Each cell is composed of different parts and the most important part of the cell is the nucleus. The nucleus has genomes and each genome is further divided into chromosomes. Normally, the numbers of chromosomes are 46 (or 23 pairs). These chromosomes are inherited from parents 23 from male and 23 from female. Each chromosome has long strands of DNA (deoxyribonucleic acid) and which are made up of numerous genes. The long strand of DNA is a double helix structure in which genes have combined or adjacent to each other. This gene sequence has four alphabets of the English language (A, G, C, T) and A is always connected with G while C will always be connected with T. This gene sequence has instructions regarding protein development. [4]

### *B. Mutation*

Cancer is the uncontrolled and abnormal growth of body cells. Instead of dying old cells and replacing them with new, new cells are created without demolishing old cells. This abnormal growth forms a mass of tissue which is called tumor. Most often the main reason for tumor and abnormal growth of cells is mutation in genetic code. Mutations in genes alter the sequence of DNA, which disturbs the original sequence of the genetic code and turns it into an abnormal cell. It can be said that mutation is the pre stage of cancer. When the mutated genes reach an excessive stage, it loses its balance and begins to develop abnormally. Mutation is occurred in the structure of DNA, since A relates to G and C always relates to T. but in case of mutation, this rule might be disturbed. [6]

## II. METHODS AND TECHNIQUES

This study proposed a solution to detect mutation in genes. It might be helpful in diagnosing cancer before it's occurrence or at initial stages. Many machine learning techniques can be useful for mutation detection after its training on gene sequence dataset. One of the main reasons of the deaths due to cancer is the late identification of the disease. For resolving this issue, many cancer detection methods have been introduced using machine learning and deep learning techniques.

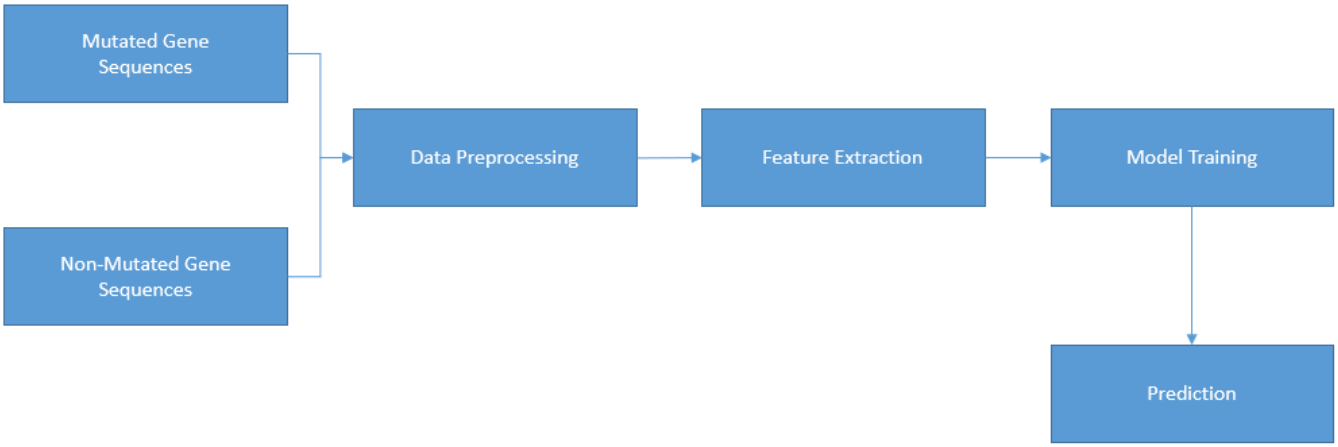


Fig. 1. Methodology

### A. Dataset

The dataset for mutation detection system is retrieved from different web portals and research sites. The non-mutated gene sequences are retrieved from NCBI [7] and mutated gene sequences of BRCA1 is retrieved HGMD [8]. Both datasets are in FASTA format and pre-processed using python script and combine them in csv format. The dataset is of BRCA1 mutated and non-mutated genes. BRCA1 is referred as breast cancer. A lot of women suffered breast cancer every year. For mutation detection to early diagnose of cancer BRCA1 gene sequences has been used.

### B. BRCA1

BRCA1 (BReast CAncer type 1) is a gene found in every woman. Originally it repairs the breaking DNA and unnecessary growth of the cells which prevents from cancer. It is also named as tumor suppressor genes. If BRCA1 damages, then it fails to avoid the alteration of genes which first lead to mutation and then ultimately to cancer. Around 12% of women are supposed to be suffered with breast cancer in their lifetime. Almost 60-65% of them suffered it due to damage of BRCA1 [9].

### C. Pre processing of Data

The dataset of mutated and non-mutated BRCA1 genetic code is downloaded from respected websites in fasta format. These sequences were in a raw form and polluted with many irregular patterns. Firstly, all unnecessary and redundant alphabets were filtered out and trimmed all the sequences to the length of 71 characters. Now, cleaned and filtered data is labeled using Python script. All mutated genes are labeled as 1 and non-mutated as 0. Then these files are merged and saved as a CSV format. The overall methodology that has been implemented is illustrated in the above fig. 1.

The above fig. 1 shows the whole procedure of implementation. After the preprocessing of dataset, final and benchmark dataset is ready for the further processes. Since, our dataset is composed of mutated and non-mutated genetic codes, labeled as '1s' and '0s'. If sequence is mutated it referred as 1 or vice versa. The next step is to extract the features from these simple genetic codes. To extract different features, we used different techniques and these features are further used for training purposes.

## III. Feature Ectraction

The technique of feature extraction for gene sequences is thoroughly different from normal text pre-processing and feature extraction. Normal text feature extraction techniques in machine learning is totally dependent on languages of text. The NLP techniques used for feature extraction or training purposes for normal text. By analyzing the previous work in this field is shows that the generic NLP techniques were not most affective in feature extraction of gene sequences. To overcome this problem, different customized methods were used. To extract the features from raw genetic codes, different mathematical and statistical approaches were used. These approaches and their brief introduction about their working is given below:

### A. Z-Curve

It's a 3D feature extraction approach. It is used in genomic sequence analysis. It generates 3 features, and it has following mathematical equation:

$$x = (\sum A + \sum G) - (\sum C + \sum T) \quad (1)$$

$$y = (\sum A + \sum C) - (\sum G + \sum T) \quad (2)$$

$$z = (\sum A + \sum T) - (\sum G + \sum C) \quad (3)$$

### B. gcContent

gcContent is the ratio of G and C in the genetic code. If gcContent rate is high, then the genetic code is more stable. The mathematical equation of gcContent is given below:

$$gc\ ratio = \frac{\sum G + \sum C}{\sum A + \sum C + \sum G + \sum T}$$

### C. atgcRatio

atgcRatio is the ratio of AT to the GC. The equation of this term is given below:

$$atgc\ ratio = \frac{\sum A + \sum T}{\sum G + \sum C} \quad (4)$$

*D. cumulativeSkew*

cumulativeSkew is the difference in count of T and G of two strands of DNA. Mathematical equation is shown below:

$$GC\ skew = \frac{\sum G - \sum C}{\sum G + \sum C} \quad (5)$$

$$AT\ skew = \frac{\sum A - \sum T}{\sum A + \sum T} \quad (6)$$

*E. pseudoKNC*

It is a Pseudo K-tuple nucleotide composition. It was originally introduced for proteins but latter it is also used for DNA. It is widely used to check the DNA sequence order losing. The equation of this phenomenon is given below:

$$\sum_{i=1}^{n} 4^i \ ;\ when\ k = n \quad (7)$$

We have used k = 3 and feature structure is X, XX, XXX in general.

*F. monoMonoKGap*

monoMonoKGap is the technique used to find the number of gaps between A, T, G and C. The equation of this technique is shown below:

$$kGap(n) = \ 4 \times 4 \times n \quad (8)$$

If n is 2, the general structure is X_X, X__X

*G. Feature Selection*

After applying the feature extraction techniques, we applied the feature selection methodology. By applying the feature extraction approaches we get bunch of features in which lots of features are not so useful and with these amounts of features the modelling and implementation of algorithms may affect. To optimize the features, we reduce the features and select only most impactful feature. We reduce the features by training the model Adaboost with default hyperparameters and select the optimized and most useful features for further processing.

## IV. Results and Discussions

*A. Estimated accuracy*

One of the most important steps during the development of the machine learning model is, the unbiased and impartial evaluation of the model. For an impartial success evaluation of implemented models, two main metrics are used to inspect the model evaluation, 1: Best and fitted testing methods should be used for score metrices. 2: Appropriate and most suitable metrices should be used to reflect the quality of model prediction.

*B. Testing via 10-fold cross validation*

Cross validation is the most common technique used to evaluate the models. This approach is very beneficial for error estimation of classifiers or models. practitioners used K-fold cross validation (KFCV) technique to check the prediction or validity of the classifiers. The KFCV, also known as rotation estimation, the benchmark or whole dataset is distributed into the K number of distinctive folds, where K is the number in which the dataset is split randomly. After dividing the entire dataset into K folds (these folds are just about equal in size), the Kth part of the data is used for predicting purpose and the remaining folds (k-1) are used to fit in the classifier for training purpose. The Prediction error is calculated for each cycle. If k=5, it's repeated for 5 times and finally predictive error for each iteration is combined as a result.

For now, K=10 and we run this iterative process on selective classifiers. The results of the training and accuracy of each model is shown in table I. We set the threshold of 85% and if any model meets the threshold, the weights of respective models are saved for prediction purposes.

Table I. KFCV Results

| **Models** | **K-folds** | **Accuracy (%)** |
|---|---|---|
| Logistic Regression | 10 | 91.44 |
| KNN | 10 | 76.43 |
| Naive Bayes | 10 | 74.47 |
| Bagging Classifier | 10 | 76.14 |
| AdaBoost Classifier | 10 | 79.39 |
| Gradient Boosting Classifier | 10 | 83.02 |
| SVM | 10 | 97.05 |
| Linear Discriminant Analysis | 10 | 89.46 |
| Extra Trees Classifier | 10 | 72.61 |

According to the results of KFCV results, we selected only three models with the highest accuracy and develop a pipeline, which is used for predictions.

*C. Formulatiion of metrics*

To evaluate the accuracy of the prediction models, we used different famous and best fit metrics. To evaluate the statistical predictions, different metrics are often used such as Accuracy, Sensitivity (Sn or ROC) and Specificity (Sp). To retrieve the outputs of these metrics first we get the confusion metric for each prediction model. The formulae of each metrics are given below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$Sn = \frac{TP}{TP + FN} \quad (10)$$

$$Sp = \frac{TN}{TN + FP} \quad (11)$$

Here, TP, FP, TN, FN shows the number of true positives, false positives, true negatives, and false negatives predictions respectively. These values could also be shown in a single table named as confusion matrix for better understanding. Table II. Shows the value of TP, FP, TN, FN for selected models.

Table II. Confusion Matrix values of selected models.

| Models | Metrices | | | | |
|---|---|---|---|---|---|
| | TP | FP | TN | FN | Accuracy (%) |
| LR | 217 | 15 | 152 | 20 | 91.33 |
| SVM | 233 | 3 | 163 | 10 | 96.81 |
| LDA | 218 | 26 | 146 | 18 | 89.21 |

In determining these values for each applied model, the above formulas are applied to check the accuracy of the prediction. By applying *(Sn)* we plotted a ROC curves for each model using iterative process. The plotted receiver operating characteristics (ROC) curve is the ratio of true positive value to the summation of true positives and false negatives. It shows the performance of the classifier with respect to all classification thresholds. It is also known as recall score of the model. The plotted curve is shown in fig. 2.

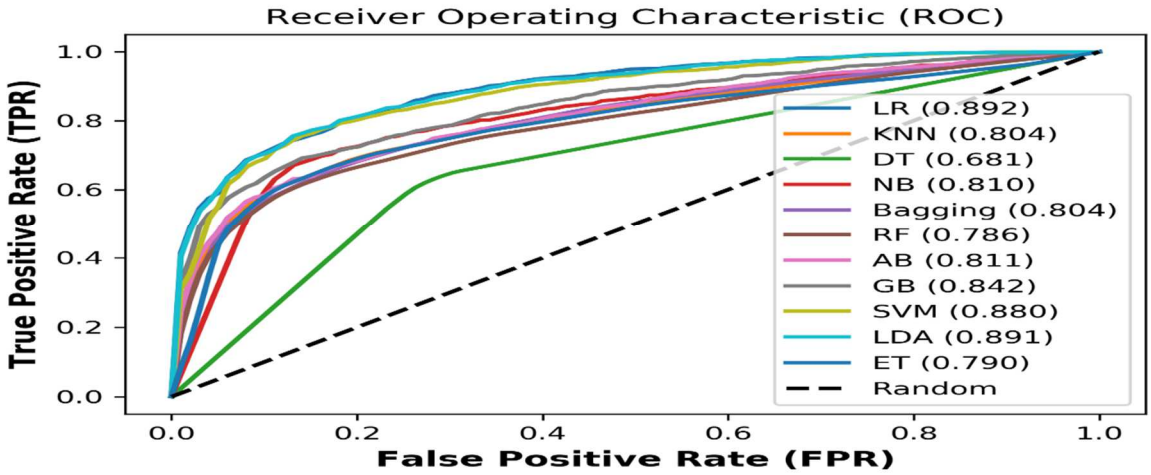


Fig. 2 ROC curves of all implemented model.

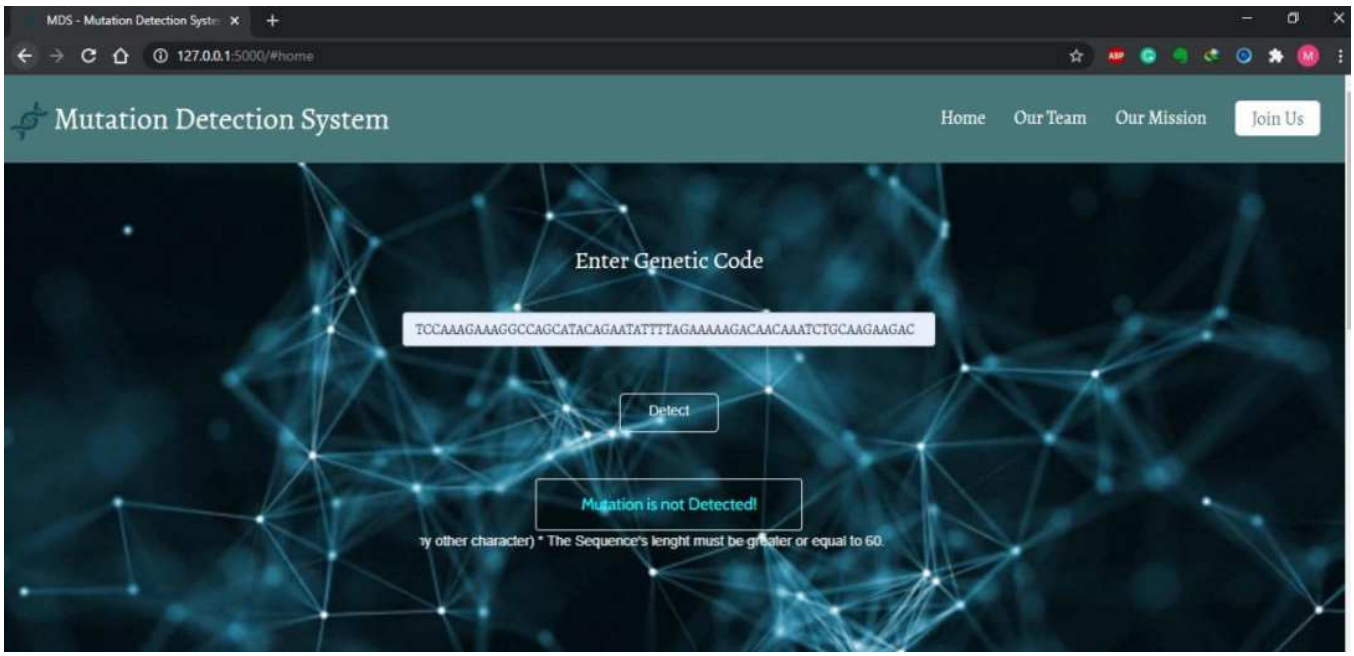


Fig. 3. The graphical user interface (GUI) of Mutation Detection System.

## V. WEB SERVER

The final and ultimate step of the whole project is the development of user-friendly interface for users. The interface is developed for the ease of the pathologists, biologists, and other users, to check the mutation in genetic code. A simple and user-friendly interface of web portal has been designed. Initially, web portal is hosted on a local server, but in future it will be available globally. The web portal is designed by following the design principles and the interface is shown in fig. 3. As it gets live on a server, The users could access this web portal globally. Following are steps should be followed to get the mutation detection results. The frontend of the mutation detection system is designed using HTML, CSS, JS and deployed using Flask (framework of python). For the better understanding, we provide a step-by-step user guide to the usage of the web portal.

When you open the web server, you will find the web page, shown in fig. 3. with the header containing four pages, **Home, Our Team**, **Our mission** and **Join us.** It's a single page web portal and the home section containing the input field for the genetic code which should be provided by the user and the button name 'Detect'. When the button is clicked, it triggers different functionalities at the backend. Firstly, it checks the validity of the input, whether it's a valid genetic code or not? It should be composed of only AGCT characters, and the length of the input string should be greater than 65. otherwise, it shows an error prompt. Once the user puts the valid input and clicks on the detect button it shows the output in the result section in the textual format.

### *A. User Model*

To demonstrate the working and role of the user, we draw a user relationship diagram with different entities of this web portal, shown in fig. 4. According to this diagram, user may belongs to different roles such as a researcher, a student, or a biologist. There are several other approaches are presented by [27-55]

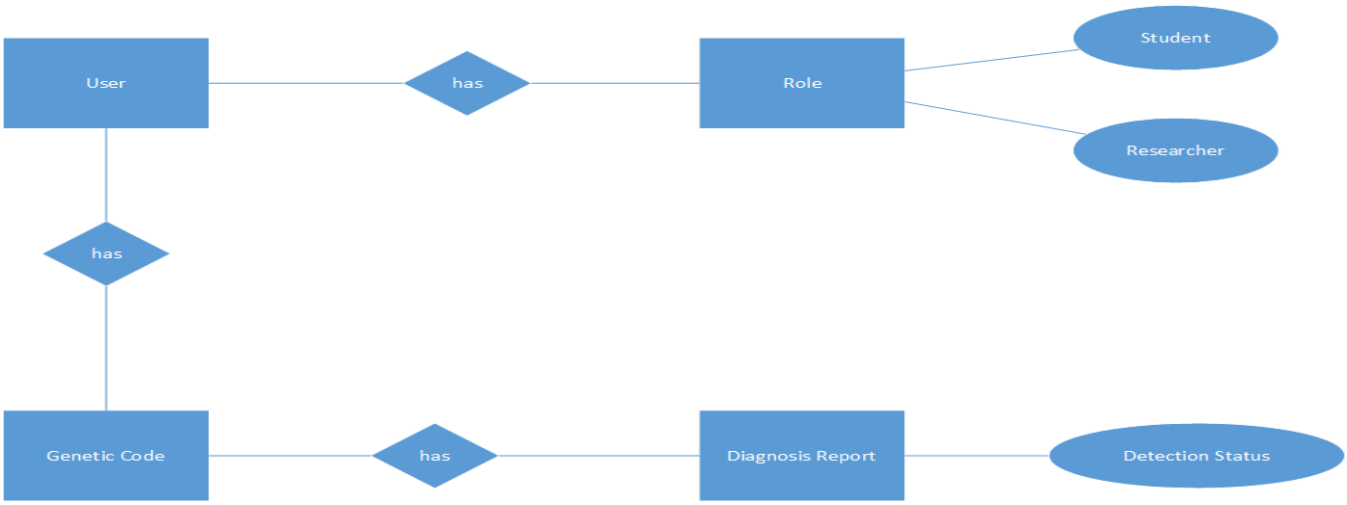


Fig. 4. User model diagram.

The above fig. 4 also shows that, user must have a human genetic code to detect the mutation in that specific sequence. Moreover, each genetic code has its own diagnosis report which also be referred as the detection status.

## VI. CONCLUSION

This paper proposed an efficient and effective model to detect mutation in genes sequence using many statistical and mathematical approaches for feature extraction. Several machine learning models were used for better results. Among all these models, SVM and LR have achieved the highest accuracy for mutation detection in gene sequence to pre-diagnose cancer. To improve the mutation detection process, different deep learning techniques could also be implemented in the future.

## REFERENCES


[1] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., Sequence-based identification of arginine amidation sites in proteins using deep representations of proteins and PseAAC. Current Bioinformatics 2020, 15 (8), 937-948.

[2] D. Dandrea, F. Soria, S. Zehetmayer, K. M. Gust, S. Korn, J. A. Witjes, and S. F. Shariat, "Diagnostic accuracy, clinical utility and influence on decision-making of a methylation urine biomarker test in the surveillance of non-muscle-invasive bladder cancer," BJU International, vol. 123, no. 6, pp. 959–967, 2019.

[3] J. Xi, A. Li, and M. Wang, "A novel network regularized matrix decomposition method to detect mutated cancer genes in tumour samples with inter-patient heterogeneity," Scientific Reports, vol. 7, no. 1, 2017.

[4] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., NPalmitoylDeep-PseAAC: A Predictor of N-Palmitoylation Sites in Proteins Using Deep Representations of Proteins and PseAAC via Modified 5-Steps Rule. Current Bioinformatics 2021, 16 (2), 294-305.

[5] Malebary, S. J.; Khan, Y. D., Evaluating machine learning methodologies for identification of cancer driver genes. Scientific reports 2021, 11 (1), 1-13.

[6] "Homo sapiens BRCA1 DNA repair associated (BRCA1), transcript variant 1 - Nucleotide - NCBI," National Center for Biotechnology Information. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore /NM_007294.3? report=fasta&to=7224. [Accessed: 03-Jan-2021].

[7] "HGMD® gene result," HGMD. [Online]. Available: http://www.hgmd.cf.ac.uk/ac/gene.php?gene=BRCA1. [Accessed: 03-Jan-2021]. 47

[8] "Homo sapiens BRCA1 DNA repair associated (BRCA1), RefSeqGene (LRG_292) - Nucleotide - NCBI," National Center for Biotechnology Information. [Online]. Available: https://www.ncbi.nlm.nih.gov/nuccore/NG_005905.2?report=fasta&to=193689. [Accessed: 03-Jan-2021].

[9] "BRCA: The Breast Cancer Gene," National Breast Cancer Foundation, 23-Oct-2020. [Online]. Available: https://www.nationalbreastcancer.org/what-is- brca. [Accessed: 05-Jan-2021].

[10] Saeed, S.; Mahmood, M. K.; Khan, Y. D., An exposition of facial expression recognition techniques. *Neural Computing and Applications* **2018,** *29* (9), 425-443.

[11] Butt, A. H.; Khan, Y. D., CanLect-Pred: A cancer therapeutics tool for prediction of target cancerlectins using experiential annotated proteomic sequences. *IEEE Access* **2019,** *8*, 9520-9531.

[12] Khan, S. A.; Khan, Y. D.; Ahmad, S.; Allehaibi, K. H., N-MyristoylG-PseAAC: sequence-based prediction of N-myristoyl glycine sites in proteins by integration of PseAAC and statistical moments. *Letters in Organic Chemistry* **2019,** *16* (3), 226-234.

[13] Amanat, S.; Ashraf, A.; Hussain, W.; Rasool, N.; Khan, Y. D., Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PseAAC. *Current Bioinformatics* **2020,** *15* (5), 396-407.

[14] Hussain, W.; Rasool, N.; Khan, Y. D., A Sequence-Based Predictor of Zika Virus Proteins Developed by Integration of PseAAC and Statistical Moments. *Combinatorial chemistry & high throughput screening* **2020,** *23* (8), 797-804.

[15] Khan, Y. D.; Alzahrani, E.; Alghamdi, W.; Ullah, M. Z., Sequence-based Identification of Allergen Proteins Developed by Integration of PseAAC and Statistical Moments via 5-Step Rule. *Current Bioinformatics* **2020,** *15* (9), 1046-1055.

[16] Mahmood, M. K.; Ehsan, A.; Khan, Y. D.; Chou, K.-C., iHyd-LysSite (EPSV): Identifying Hydroxylysine Sites in Protein Using Statistical Formulation by Extracting Enhanced Position and Sequence Variant Feature Technique. *Current Genomics* **2020,** *21* (7), 536-545.

[17] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., IPhosS (Deep)-PseAAC: Identify phosphoserine sites in proteins using deep learning on general pseudo amino acid compositions via modified 5-Steps rule. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2020**.

[18] Shah, A. A.; Khan, Y. D., Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification. *Scientific Reports* **2020,** *10* (1), 1-10.

[19] Almagrabi, A. O.; Khan, Y. D.; Khan, S. A., iPhosD-PseAAC: Identification of phosphoaspartate sites in proteins using statistical moments and PseAAC. *Biocell* **2021,** *45* (5), 1287.

[20] Awais, M.; Hussain, W.; Rasool, N.; Khan, Y. D., iTSP-PseAAC: Identifying Tumor Suppressor Proteins by Using Fully Connected Neural Network and PseAAC. *Current Bioinformatics* **2021,** *16* (5), 700-709.

[21] Nadeem, M. W., Goh, H. G., Ponnusamy, V., Andonovic, I., Khan, M. A., & Hussain, M. (2021, October). A fusion-based machine learning approach for the prediction of the onset of diabetes. In Healthcare (Vol. 9, No. 10, p. 1393). Multidisciplinary Digital Publishing Institute.

[22] Awan, M. J., Farooq, U., Babar, H. M. A., Yasin, A., Nobanee, H., Hussain, M., ... & Zain, A. M. (2021). Real-time DDoS attack detection system using big data approach. Sustainability, 13(19), 10743.

[23] T. M. Ghazal, M. Anam, M. K. Hasan, M. Hussain, M. S. Farooq et al., "Hep-pred: hepatitis c staging prediction using fine gaussian svm," Computers, Materials & Continua, vol. 69, no.1, pp. 191–203, 2021.

[24] Khan, A. H., Hussain, M., & Malik, M. K. (2021). Arrhythmia Classification Techniques Using Deep Neural Network. Complexity, 2021.

[25] Khan, A. H., Hussain, M., & Malik, M. K. (2021). Cardiac disorder classification by electrocardiogram sensing using deep neural network. Complexity, 2021.

[26] M. W. Nadeem, H. G. Goh, M. A. Khan, M. Hussain, M. F. Mushtaq et al., "Fusion-based machine learning architecture for heart disease prediction," Computers, Materials & Continua, vol. 67, no.2, pp. 2481–2496, 2021.

[27] Khan, A. H., Hussain, M., & Malik, M. K. (2021). ECG Images dataset of Cardiac and COVID-19 Patients. Data in Brief, 106762.

[28] M. Anam, V. A. Ponnusamy, M. Hussain, M. W. Nadeem, M. Javed et al., "Osteoporosis prediction for trabecular bone using machine learning: a review," Computers, Materials & Continua, vol. 67, no.1, pp. 89–105, 2021.

[29] M. M. Ahmed, S. A. Shehri, J. U. Arshed, M. U. Hassan, M. Hussain et al., "A weighted spatially constrained finite mixture model for image segmentation," Computers, Materials & Continua, vol. 67, no.1, pp. 171–185, 2021.

[30] Nadeem, M. W., Goh, H. G., Ali, A., Hussain, M., & Khan, M. A. (2020). Bone Age Assessment Empowered with Deep Learning: A Survey, Open Research Challenges and Future Directions. Diagnostics, 10(10), 781.

[31] Khalid, H., Hussain, M., Al Ghamdi, M. A., Khalid, T., Khalid, K., Khan, M. A., ... & Ahmed, A. (2020). A Comparative Systematic Literature Review on Knee Bone Reports from MRI, X-rays and CT Scans Using Deep Learning and Machine Learning Methodologies. Diagnostics, 10(8), 518.

[32] Malik, H., Farooq, M. S., Khelifi, A., Abid, A., Qureshi, J. N., & Hussain, M. (2020). A Comparison of Transfer Learning Performance Versus Health Experts in Disease Diagnosis From Medical Imaging. IEEE Access, 8, 139367-139386.

[33] Rehman, A. U., Hussain, M., Idress, M., Munawar, A., Attique, M., Anwar, F., & Ahmad, M. (2020). E-cultivation using the IoT with Adafruit cloud.

[34] Nadeem, M. W., Ghamdi, M. A. A., Hussain, M., Khan, M. A., Khan, K. M., Almotiri, S. H., & Butt, S. A. (2020). Brain tumor analysis empowered with deep learning: A review, taxonomy, and future challenges. Brain sciences, 10(2), 118.

[35] Manzoor, A., Hussain, M., & Mehrban, S. (2020). Performance Analysis and Route Optimization: Redistribution between EIGRP, OSPF & BGP Routing Protocols. Computer Standards & Interfaces, 68, 103391.

[36] Mehrban, S., Nadeem, M. W., Hussain, M., Ahmed, M. M., Hakeem, O., Saqib, S., ... & Khan, M. A. (2020). Towards secure FinTech: A survey, taxonomy, and open research challenges. IEEE Access, 8, 23391-23406.

[37] Abid, A., Manzoor, M. F., Farooq, M. S., Farooq, U., & Hussain, M. (2020). Challenges and Issues of Resource Allocation Techniques in Cloud Computing. KSII Transactions on Internet and Information Systems (TIIS), 14(7), 2815-2839.

[38] Faheem, M. R., Anees, T., & Hussain, M. (2019). The Web of Things: Findability Taxonomy and Challenges. IEEE Access, 7, 185028-185041.

[39] Nadeem, M. W., Goh, H. G., Hussain, M., Hussain, M., & Khan, M. A. (2021). Internet of Things for Green Building Management: A Survey. In Role of IoT in Green Energy Systems (pp. 156-170). IGI Global.

[40] Iqbal, S., Hussain, M., Munir, M. U., Hussain, Z., Mehrban, S., & Ashraf, M. A. (2021). Crypto-Currency: Future of FinTech. In Research Anthology on Blockchain Technology in Business, Healthcare, Education, and Government (pp. 1915-1924). IGI Global.

[41] Hussain, M., Nadeem, M. W., Iqbal, S., Mehrban, S., Fatima, S. N., Hakeem, O., & Mustafa, G. (2021). Security and Privacy in FinTech: A Policy Enforcement Framework. In Research Anthology on Concepts, Applications, and Challenges of FinTech (pp. 372-384). IGI Global.

[42] Khan, A. G., Zahid, A. H., Hussain, M., & Riaz, U. (2019, November). Security Of Cryptocurrency Using Hardware Wallet And QR Code. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-10). IEEE.

[43] Khan, A. G., Zahid, A. H., Hussain, M., Farooq, M., Riaz, U., & Alam, T. M. (2019, November). A journey of WEB and Blockchain towards the Industry 4.0: An Overview. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-7). IEEE.

[44] Hussain M, Javed W, Hakeem O, Yousafzai A, Younas A, Awan MJ, Nobanee H, Zain AM. Blockchain-Based IoT Devices in Supply Chain Management: A Systematic Literature Review. Sustainability. 2021; 13(24):13646.

[45] Nadeem, M. W., Hussain, M., Khan, M. A., Munir, M. U., & Mehrban, S. (2019, November). Fuzzy-Based Model to Evaluate City Centric Parameters for Smart City. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-7). IEEE.

[46] Zainab, M., Usmani, A. R., Mehrban, S., & Hussain, M. (2019, November). Fpga based implementations of rnn and cnn: A brief analysis. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-8). IEEE.

[47] Hassan, M., Hussain, M., & Irfan, M. (2019, November). A Policy Recommendations Framework To Resolve Global Software Development Issues. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-10). IEEE.

[48] Rafique, I., Fatima, K., Dastagir, A., Mahmood, S., & Hussain, M. (2019, November). Autism Identification and Learning Through Motor Gesture Patterns. In 2019 International Conference on Innovative Computing (ICIC) (pp. 1-7). IEEE.

[49] Nadeem, M. W., Hussain, M., Khan, M. A., & Awan, S. M. (2019, July). Analysis of Smart Citizens: A Fuzzy Based Approach. In 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) (pp. 1-5). IEEE.

[50] Hussain, W.; Rasool, N.; Khan, Y. D., Insights into Machine Learning-based approaches for Virtual Screening in Drug Discovery: Existing strategies and streamlining through FP-CADD. *Current Drug Discovery Technologies* **2021,** *18* (4), 463-472.

[51] Khan, Y. D.; Khan, N. S.; Naseer, S.; Butt, A. H., iSUMOK-PseAAC: prediction of lysine sumoylation sites using statistical moments and Chou's PseAAC. *PeerJ* **2021,** *9*, e11581.

[52] Malebary, S. J.; Khan, R.; Khan, Y. D., ProtoPred: Advancing Oncological Research Through Identification of Proto-Oncogene Proteins. *IEEE Access* **2021,** *9*, 68788-6.

[53] Malebary, S. J.; Khan, Y. D., Identification of Antimicrobial Peptides Using Chou's 5 Step Rule. *CMC-COMPUTERS MATERIALS & CONTINUA* **2021,** *67* (3), 2863-2881.

[54] Naseer, S.; Ali, R. F.; Khan, Y. D.; Dominic, P., iGluK-Deep: computational identification of lysine glutarylation sites using deep neural networks with general pseudo amino acid compositions. *Journal of Biomolecular Structure and Dynamics* **2021**, 1-14.19.

[55] Naseer, S.; Hussain, W.; Khan, Y. D.; Rasool, N., Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. *Analytical Biochemistry* **2021,** *615*, 114069.