Cancer Type Driver Classification Accuracy Using Spark ML Technology

Daniel Mago Vistro[1], Muhammad Shoaib Farooq[2], Attique Ur Rehman[3], and Hafiz Abdullah Tanveer[4]
[1]School of Computing, Asia Pacific University, Kuala lumpur, Malaysia.
[2,3,4]School of System and Technology, University of Management and Technology, Pakistan.
[1]Daniel.mago@staffemail.apu.edu.my, [2]shoaib.farooq@umt.edu.pk, [3]F2019288002@umt.edu.pk,
[4]S2020114016@umt.edu.pk

Received: date; Accepted: date; Published: date
Corresponding author email: daniel.mago@staffemail.apu.edu.my

Abstract— **In this paper, analysis of genes extracted from the body has been performed that can be a driver of tumor, resulting in a cancer of different types like breast cancer etc. motived by the BIGBIOCL. Classifier with Alternative and Multiple Rule Based (CAMUR) is a core algorithm that is applied here to dissect large datasets. For the purpose to acquire the desire goal, Apache Spark as well as MLlib is used, on stack of Hadoop in local mode. The practice has been performed using the decision tree as well as random forest separately. As far as the deployed data is concerned, in terms of measurement of F and efficiency, random forest has shown the better results. For the objective of extraction of genes and other pertinent models, deletion of features has been performed with the deployment of iterative algorithm as proposed earlier CAMUR with modified version. Finally, the extracted results are facilitated to biologist, so they can analyzed the extraction is related or either can be a driver of cancer.**

Keywords— **Tumor; genes; CAMUR; Hadoop; Apache Spark; BIGBIOL; MLlib; Thyroid cancer**

1. Introduction

As the world is growing day by day, there are some diseases that is generating as well. And at the same time the cure of these diseases is being generated by the researchers by using the state of the art technologies. But there are some disease that are becoming endemic. Cancer is one those diseases that is endemic, that is not contagious at all. But the patient with this disease has very less chances to cure with this disease. Especially, cancer has four stages, and if the person is in third or fourth he is more likely to succumb by this. The core reason behind the cancer is the tumor. When eukaryotic cell will be uncontrolled division and destroyed the other cells and the surrounding tissues the tumor will increased this is known as cancer. One six death are occurred by cancer according to the report of world health organization. There will be no modern treatment of cancer in human race that can be fully understand its procedure to cure. It can cause by the outer agents like bacteria, germs, virus, chemical, etc. it can affect the DNA and cause human cancer. Hadoop process can find the cancer and given the solution to the doctors that can help him in treatment. Breast cancer is a type of cancer find in women. By the helping of hadoop big data can find the treatment that can be help full for doctors.

2. Literature Review

DNA Methylation is one of the most strongly contemplated hereditary change in warm blooded animals including repeatable combination adjustments of nucleotides relevant to the respective DNA [1]. Specifically, the conversion is possible dissolving the catalyst which belongs to methyl alliance, and this will help in transforming to cytosine version 5 from cytosine, the transformed material also belongs to methyl group. Traditionally in simple scenario, transformation will bring enticing characteristics, resulting in evidence of accurate visualization and regularization of genes [2]. It is a fact that there are around 30 hundreds of thousands cytosine in a normal human [3]. As methylation gets hyper in the area of gene, and other genes can be deactivated by the utilization of methylation in every branch of a cancer. So for this fact it is a major cause behind producing genes that will suppress the driver tumor. All this happened just because of cell state conversion. Hypo methylation which is also considered as globally is behind the cause that disturbs the equilibrium of genomic [4]. As it is clear from the above facts that, behind the inactivation there is a group called methylation. So it is concluded that the analyzing these genes belong to methylation is quite challenging. Beside of these problems there is another problem which is associated with computational cost as implementation of sequential models requires it [5]. As it is a common problem that the larger datasets are susceptible to classifier, they tend to show error so it is addressed [6].

Particularly for this, we center on appropriation of enormous information innovations for the use of arrangement calculations on huge datasets of methylation. Regardless of whether there are a wide range of definitions of huge information, "Large information alludes to datasets whose size is past the capacity of run of the mill database programming devices to catch, store, oversee also, and investigate" [7, 20, 21]. This definition doesn't concentrate on explicit information size, yet on the innovation we receive to deal with those datasets. So as the previous studies suggest that use of the large dataset to find a driver of tumor so they can be detected earlier and treated. These drivers of genes is detected and extracted by the help of various supervised machine learning algorithm on large datasets. As proposes in previous studies there is famous algorithm SVM [8,9] which is deployed and there are still other algorithms like Naïve Bayes [10] that can be used to produce the same action as well. A past characterization concentrate on DNA methylation [11] proposed MethPed, a device for the distinguishing proof of pediatric tumors. Specialists constructed the grouping model behind MethPed from DNA methylation datasets with 450

thousand of highlights. They right off the bat applied countless relapse calculations to choose a subset of highlights with the most elevated prescient force; at that point, they received Random Forests to construct the grouping model. On the opposite, we need to apply grouping calculations to the whole dataset so as to get an enormous number of CpG destinations and their related genomic areas. Another examination [12, 19] depicted methylKit, a R bundle for the investigation of DNA methylation information. This bundle receives a solo AI approach, working on unlabeled information. methylKit works in-memory and, regardless of whether it is multi-strung, its execution is restricted to a solitary machine. On our side, we need to perform administered AI on a group of computational hubs, so as to have the option to scale with the expanding measurement of information.

The calculation proposed in our investigation is roused by CAMUR for being applied to enormous information datasets. CAMUR (Classifier with elective and numerous standard based models) is an order strategy that iteratively registers a standard based characterization model, disposes of from the info dataset blends of separated highlights, furthermore, rehashes the arrangement until a halting condition is confirmed [13]. The consequence of a CAMUR calculation is a lot of arrangement models. CAMUR chipped away at RNA sequencing disease datasets with around 20 thousand highlights. In this work, we plan and create BIGBIOCL, a various tree-based classifier, to break down DNA methylation datasets with in excess of 450 thousand highlights [14]. Our objective is to remove applicant methylated destinations and their related qualities in barely any hours. Tumors of the focal sensory system (CNS) are the most normal strong malignancies in kids, speaking to around 20 % of all youth disease cases [15]. By and large endurance of kids with cerebrum tumors is around 70 % be that as it may, changes exceptionally relying upon type and area of the tumor. Order of pediatric tumors into natural pertinent elements is testing and imperatively significant in deciding the proper treatment convention for a particular patient [2, 3]. Youth disease survivors regularly experience considerable long haul reactions from the treatment. Picking the correct treatment and staying away from superfluous treatment is in this manner significant. A suitable reproducible classifier is in this manner earnestly expected to characterize great what's more, poor treatment reaction subgroups and for the assessment of results got from clinical preliminaries all together to approve the strength of new medications explicitly planned to specifically influence sub-atomic focuses in the separate subclasses. The most widely recognized clinical conclusion bunches incorporate pilocytic astrocytoma, high-grade glioma/glioblastoma (GBM), diffuse natural pontine glioma (DIPG), ependymoma, and crude neuroectodermal tumor of the (CNS-PNET), medulloblastoma (cerebellar PNET), what's more, supratentorial PNET (sPNET); be that as it may, there are in excess of 100 distinctive histological subtypes. Utilizing ordinary parameters, for example, area and histology (WHO models) for finding won't catch the full image of these tumors and consequently lead to both underhand overtreatment just as hamper the recognizable proof of prognostic components and atomic biomarkers [16]. Past investigations have indicated that methylation profiling utilizing the Illumina 450K methylation exhibits can isolate a few pediatric cerebrum tumor analyze including the four medulloblastoma subgroups; sonic hedgehog (MB_SHH), WNT (MB_WNT), bunch 3 (MB_Gr3), and gathering 4 (MB_Gr4) [5–9]. In any case, an order instrument for diagnosing an obscure tumor is as yet deficient. In the current study, we built up a characterization device, MethPed, which can vigorously distinguish mind tumor judgments and subgroups utilizing genome-wide DNA methylation cluster information, which beats past strategies utilizing for instance quality articulation information [17].

3. **Hadoop**

Hadoop is basically big data frame work. It allows to store and process large data sets in parallel and distributed fashion. Hadoop helps to store large amount of data in its HDFS storage parts in rows and columns.it will be interconnect with each file where the data will be stored. In Hadoop Map Reduce in Hadoop the work into different paths and complete the process faster. For Example, the teacher given the work to the four student to find the word Storage from chapter one. The students start on working. Three student answer that the word Storage repeat 35 times in this chapter and one asked the word Storage repeat 36 time. The teacher proceed the answer of the majority Students. Hats the working of Map reduce. The process of master slave of Architect is to distribute the terms into other persons. For Example:-The project manager have the project it can given to the four persons then suddenly the one person will be ill then that work will be delay. Then the project manager use the master slave of Architect to distribution the work into three persons by one front roll work and second work will be on back end. HDFS Hadoop distributed File system is only to stored data into files in Hadoop. HDFS core component basically the storage of Hadoop in the process. Name Node, Data Node and secondary Name Node.it is the process of storage of Hadoop files system.

4. **Apache Spark**

As the core concern in any of the Big Data project is that, we have to analyze large data sets require higher computing power, resulting in a more waiting time. As mentioned in a literature review that the past work in big data has been performed majorly on Hadoop. The Hadoop provides higher scalability and flexibility as it also requires higher computing but the major issue was that it provides higher waiting time. In real time big data projects, there should be no latency between running and post a query. So Apache Spark is new framework that will help in addressing these problems as this will help in speeding up the computational power. It is a misconception that the Spark is an advanced version of the Hadoop, but it is not the case as it can be implemented in several ways, and Hadoop is one of the ways to implement it. It is a prove that it owns cluster management tool. So simply there are two ways a spark can utilize Hadoop. One way is to use it as storage and the second way is to implement it as processing. But the spark only uses hadoop to implement the storing process in a more flexible way. As explained earlier it is based on the hadoop, so it actually extends all the methods and algorithms of hadoop in a more efficient way, and to speed up the computation process soit can analyze

larger datasets in more quick way. In our case we have used Apache spark on front of Hadoop, especially in a local mode.

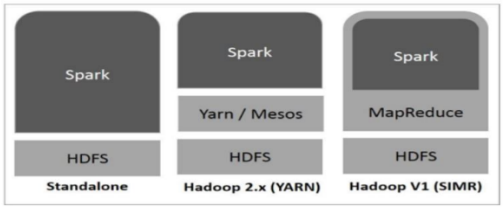

Figure 1: Spark architecture.

As it is shown above that how in three different spark can be used on top of Hadoop with an efficient way and scalable way.

5. Random Forest

Random Forest is a grouped classifier made using many decision trees models. Ensemble models combine the results from different models. Random Forest is a versatile algorithm capable of performing both, regression as well as classification. It is commonly used in predictive modeling and machine learning techniques.

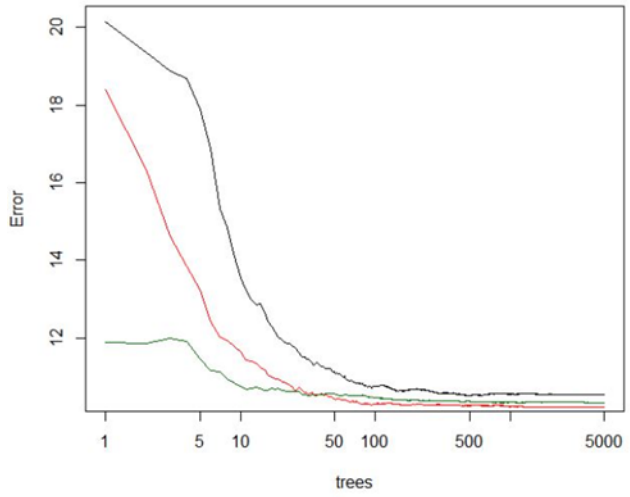

Figure 2: Trees Performance Capabilities.

The working principle of random forest is as follow; it actually randomly selects features from data, calculates the best split point among the features for node, and splits the node into two daughter nodes using the best split. Repeat first three steps until (N) number of Nodes. Finally build your Forest by repeating steps 1-4 for (D) number of times. There are several benefits of random forest which include the performance is high in comparison with all other supervise learning algorithm.

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} W_j(x_i, x')\, y_i = \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} W_j(x_i, x')\right) y_i.$$

As feature engineering is one of the most important in any of the machine learning pipeline. So random forest gives the best analyzer for the feature that gives the best insights out of the data, and these features can provide highest accuracy and the precision on a dataset. In fact, cross validation is performed with training a model, by calculating the cost on the specific input and after that test error is evaluated. But in case of random forest, without calculating the cost associated with a specific Training, test error is calculated. As far as the disadvantages are concerned, the only disadvantage is the computational cost becomes higher as the depth gets higher.

As it is explained earlier the working principle of random forest, the individual weight is average of several trees that is considered to be as N branches as equation mentioned below.

6. Decision Trees

Decision tree is another algorithm with greater efficiency as well as performance. This algorithm is used for the purpose of both classification and regression. As name depicts that it is like a tree moves from upside to downward as root on its top. The main working principle behind this algorithm is splitting.

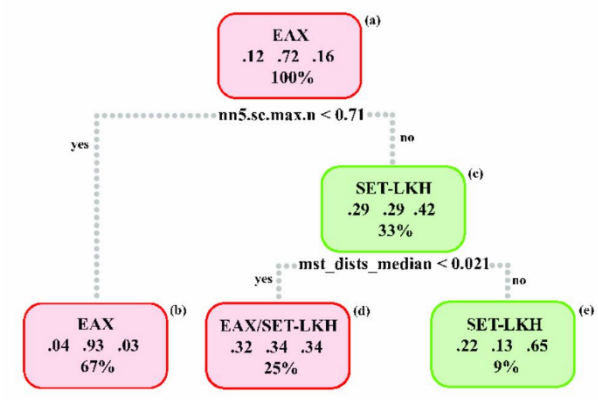

Figure 3: Decision Tree depth

The need of huge amount of data like genomic data is growing, that's why the techniques like next generation sequencing are also growing. For the same purpose it is desired to have a knowledge that can discover models from these sequences. So for that classifier with alternative multiple rule base model, as the name suggests the model will actually extracts various models, having various modification, and equivalency on their performance. The working principle of CAMUR is that it actually performs iterations for the various models, evaluates the combination set of the attributes presented in a dataset, and eliminates these power sets. It would not stop until the defined input parameters are not satisfied.

One of the best things about CAMUR is that it includes the tools used for queries also has the database repository. As far as our project is concerned, this project used three datasets of breast, kidney and thyroid and the data set is processed and deployed.

7. DATA-SET

The data set that was used in this project is quite complicated. As the cancer genome atlas was a main source from where the methylation DNA has been extracted, and the respective BIGBIOCL has been deployed on this dataset. Different types of cancer specifically Breast cancer which is medically named as breast invasive carcinoma, thyroid cancer which is named as thyroid carcinoma, and finally the kidney cancer which is medically named as kidney renal papillary cell carcinoma are described in this comma separated file.

Dataset	Number of samples	Number of features
BRCA	897	485,512
THCA	571	485,512
KIRP	321	485,512

Figure 4: Deployed dataset.

The comma separated file has header which must be skipped the data starts from the second row. Also the column must be removed which is numbered as first column, that actually contains the tissues having different codes, final the remaining columns are the features. Similarly, the last column is a target variable, which contains gene that will be the core reason to produce as tumor or it is normal. The features in a data have double or floating values.

The below table shows the final structure that has been retrieved for the thyroid cancer.

Sample ID	cg13869341	...	cg00381604	Class
TCGA-A7-A0DC-11	0.971644	...	0.017485	Tumoral
TCGA-BH-A0BV-11A	0.925557	...	?	Normal
TCGA-BH-A0DZ-11A	0.907020	...	0.019204	Tumoral

Figure 5: Extracted Dataset.

In our case, we have used the beta value. The core purpose of using the beta value is to entrench the level of DNA methylation. As it would be very useful in defining the patterns. It is clear from the below equation that it is actually the ratio of intensities. Overall methylation intensity over methylation intensity of allele.

$$\beta_n = \frac{\max(Meth_n, 0)}{\max(Meth_n, 0) + \max(Unmeth_n, 0) + \varepsilon}$$

Input Parameters	Description
Max Number Of Iterations	Stopping condition in case min F Measure is not reached
Min F-Measure	Main Stopping Condition
Max Depth	Random Forest Parameter
Max Bins	Random Forest Parameter
Number Of Trees	Splitting principle

Figure 6: Input Parameters

8. Data Preprocessing

As, it is the vital in any AI project's pipeline as in this case the output variable or a target variable is just a string value, similarly there are a lot of null value in the a target variables. At a same time there are thousands of rows for features that contain the double or floating values, as well as these rows also contain the vacant cells that will be substantial to bring the accuracy and precision of a system to much lower value. All these issues must be addressed by the data preprocessing. As the first issue the target variables are converted into 0 and 1. The normal is encoded as 0 and the tumor is encoded as 1. Similarly, all the double or floating values are also normalize for the purpose of accurately deployment of the features, so these are trained properly. Moreover, the vacant cells in feature columns are translated as question mark that will make the machine to learn the value there is not such condition.

9. Proposed Technique

Classifier with Alternative and Multiple Rule Based (CAMUR) is a core algorithm that is applied here to dissect large datasets. For the purpose to acquire the desire goal, Apache Spark as well as MLlib is used, on stack of Hadoop in local mode. The practice has been performed using the decision tree as well as random forest separately. As far as the deployed data is concerned, in terms of measurement of F and efficiency, random forest has shown the better results. For the objective of extraction of genes and other pertinent models, deletion of features has been performed with the deployment of iterative algorithm as proposed earlier CAMUR with modified version. All the results and execution has been discussed in results and discussion session.

10. Results

So this section is pertinent to outputs acquired by the execution of proposed algorithm. The following table shows the settings for ransom forest, in an earlier mentioned mode for spark.

ID	Memory	Threads	Max depth	Max bins	#Trees	Impurity
7	5 GB	7	5	16	5	Gini
8	12 GB	7	5	16	5	Gini
9	12 GB	7	5	16	10	Gini

Figure 7: Random forest setting in local mode

Similarly, the results obtained from the execution of random forest on the above tuning parameters.

ID	Build time	Evaluation time	F-Measure	#Features
7	1 h 35 min	20.37 min	98.92%	33
8	25.53 min	1.73 min	98.47%	40
9	28.87 min	1.97 min	98.83%	77

Figure 8: Outputs on Settings

As it can be seen from the above table, that the algorithms almost learnt in less than one hour, that is highly appreciable. There are some other assumption that can be inferred from the above table is that as it is able to retrieve only few attributes that can be of highly beneficial that it will save time by just retrieving those features that is of 100% of unique importance. At the same time, parallelization has less chance to contribute with decision tree. And this would be lucrative especially when there is need of multiple process and iterative.

ID	Memory	Threads	Max depth	Max Bins	#Trees	Stopping condition
10	18 GB	7	5	16	5	*F*-measure <98%
11	18 GB	7	5	16	10	*F*-measure <98%
12	18 GB	7	5	16	10	*F*-measure <97%
13	18 GB	7	5	16	20	*F*-measure <99%

Figure 9: Setting of BIGBIOCL in local mode

The above and below table shows that the running of spark in local mode. As it is explained earlier the feeding data is of breast cancer. 224 genes has been retrieved approximately in just 58 minutes by only two iterations, it is possible by setting the F measure equal to 99% by BIGBIOCL.

ID	Overall time	#Iterations	#Features	#Distinct genes
10	3 h 33 min	8	331	230
11	13 h 16 min	26	2345	1460
12	46 h 34 min	96	9780	5072
13	1 h 2 min	2	329	224

Figure 10: BIGBIOCL Output

11. Discussions

As it is shown in the experiments that BIGBIOCL has performed better for the models that have attributes in thousands of instances, as it is only made possible just by the cause of software that run on front of hadoop, that actually brought time elapsed down by the value of 75% with working of spark in the earlier mode. As in many cases the structure as well as the size of the input dataset increases, resulting in a need of or it is more obvious to say that demanding more nodes to add in its own cluster. It is only possible by the addition of parallelism to the respective software. So to move in this orientation there are number of parameter that must be tuned in order to enhance parallelism, that is all about above algorithms. The most important point to notice here is that training time increases as we increase the depth as well as the number of trees, resulting in increase of the computational cost, and the cluster has more nodes in it.

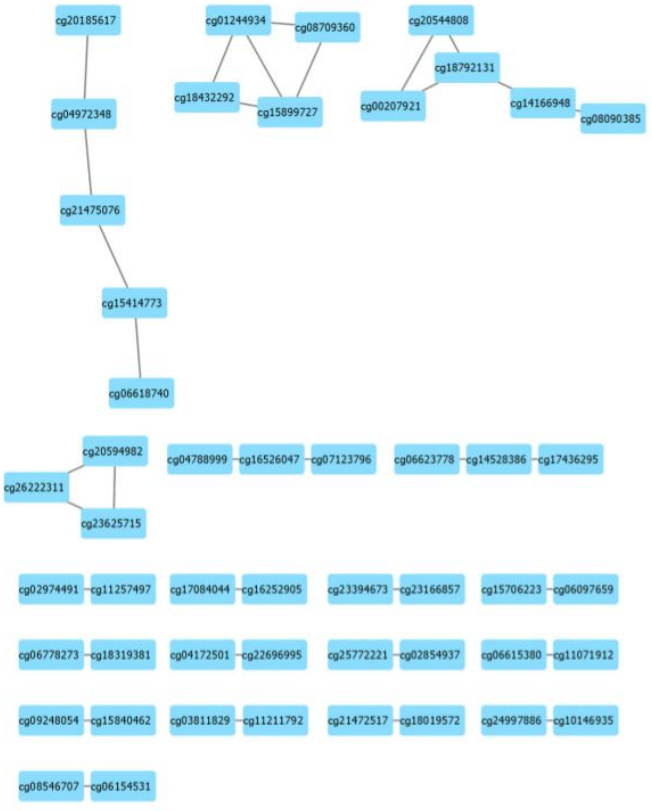

Figure 11: Brest cancer correlation output

New knowledge is revealed when comparison is made of BIGBIOCL with other algorithms such as SVM and probability theorem in context of retrieved genes. Also a comparison is made with DNA methylation. The best part about BIGBIOCL is that actually it performed on all dataset comprehensively means on all the features without deletion of any, even the features are in millions. And the motivation behind this magic is taking along the big data technology. As dataset is taken using the novel platform named as methylation450. As it is a fact that there are millions of genomes in human, even though this platform able to translate about approximately 98% of this. So this would be very helpful to recognize the bringers of cancer. So in this proposed work, the resulting proposed methodology will help the future scientist to analyze the retrieved genes that can bring the tumor. Novel paradigm will be that it also would helpful in analyzing the anonymously larger dataset. Let's assume that there are approximately twenty five thousands genes in a body, so it would be lucrative for the scientist to gaze at only limited amount of instances, so they can infer the best out of it. For example, as the suppressor of genes for PIK3CA is mostly related to the cancer of breast. An this is main lead that bring this type of cancer.

The output breast cancer correlation is shown in figure. Similarly, the correlation for the kidney as well as thyroid is shown in figure below.

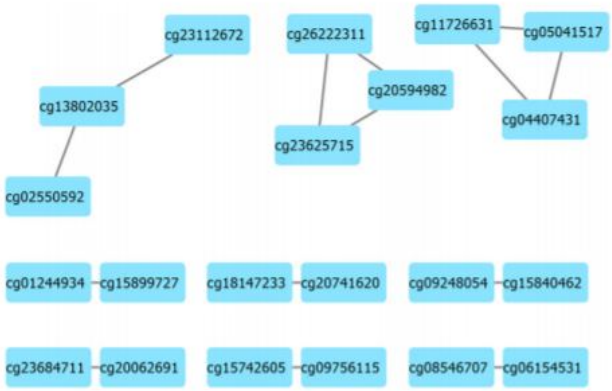

Figure 12: Thyroid cancer correlation output

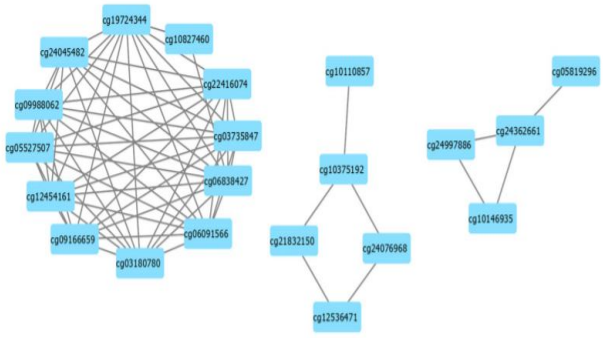

Figure 13: Kidney cancer correlation output

12. Conclusion

It is concluded that, analysis of genes extracted from the body has been performed that can be a driver of tumor, resulting in a cancer of different types like breast cancer etc. Classifier with

Alternative and Multiple Rule Based (CAMUR) is a core algorithm that is applied here to dissect large datasets. 70 % data is used as input and remaining 30% of data is used as test data. For the purpose to acquire the desire goal, Apache Spark as well as MLlib is used, on stack of Hadoop in local mode. The practice has been performed using the decision tree as well as random forest separately. As far as the deployed data is concerned, in terms of measurement of F and efficiency, random forest has shown the better results. For the objective of extraction of genes and other pertinent models, deletion of features has been performed with the deployment of iterative algorithm as proposed earlier CAMUR with modified version. Finally, the extracted results are facilitated to biologist, so they can analyzed the extraction is related or either can be a driver of cancer.

References

[1] A. Akalin, et al., methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles, Genome Biol. 13 (2012) R87.

[2] I. Antonucci, et al., A new case of “de novo” BRCA1 mutation in a patient with early-onset breast cancer, Case Rep. Clin. 5 (3) (2017) 238–240.

[3] T.E. Bartlett, et al., A DNA methylation network interaction measure, and detection of network oncomarkers, PLoS ONE 9 (1) (2014) e84573.

[4] S.B. Baylin, et al., Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer, Hum. Mol. Genet. 10.7 (2001) 687–692.

[5] Hassan, B., Farooq, M. S., Abid, A., & Sabir, N. (2016). Pakistan Sign Language: computer vision analysis & recommendations. VFAST Transactions on Software Engineering, 4(1), 1-6.

[6] Aziz, O., Farooq, M. S., Abid, A., Saher, R., & Aslam, N. (2020). Research trends in enterprise service bus (ESB) applications: a systematic mapping study. IEEE Access, 8, 31180-31197.

[7] Hajjo, R. (2018, November). then Ethical Challenges of Applying Machine Learning , Artificial intelligence on Cancer Care. on 2018 1st international Conference on Cancer Care and alsomatics (CCI) (pp. 231-231). IEEE

[8] Abujamous, L., Tbakhi, A., Odeh, M., Alsmadi, O., Kharbat, F. F., & Abdel-Razeq, H. (2018, November). Towards Digital Cancer Genetic Counseling. on 2018 1st international Conference on Cancer Care and alsomatics (CCI) (pp. 188-194). IEEE

[9] Partanen, V. M., Anttila, A., Heonävaara, S., Pankakoski, M., Sarkeala, T., Bzhalava, Z., ... & Ágústsson, Á. I. (2019). Screen–an interactive tool and also presenting cervical cancer screening indicatand on the countries. Acta Oncologica, 58(9), 1199-1204.

[10] Chen, S. B., & Hsu, C. Y. (2011, August). The TCR cancer registry repository for annotating cancer data. In 2011 2nd IEEE International Conference on Emergency Management and Management Sciences (pp. 297-300). IEEE.

[11] Daabes, A. S. A. (2018, November). Cancer treatment using herbals on Arabic social media: content analysis of YouTube videos. on 2018 1st international Conference on Cancer Care and alsomatics (CCI) (pp. 215-216). IEEE.

[12] Zhu, L., Han, B., Li, L., Xu, S., Mou, H., & Zheng, Z. (2009, October). Null Space LDA Based Feature Extraction of Mass Spectrometry Data fand also Cancer Classification. on 2009 2nd international Conference on Biomedical Engineering , and alsomatics (pp. 1-4). IEEE.

[13] Honkal, G. W., Farrell, D., Hook, S. S., Panaro, N. J., Ptak, K., & Grodzonski, P. (2011, August). Engoneerong a change on cancer diagnosis , therapy through nanotechnology. on 2011 11th IEEE international Conference on Nanotechnology (pp. 825-829). IEEE.

[14] Jugessur, A. S Textand, A. Grierson (2018, December). Nanometer Scale Coating Using Atomic Les. on 2018 IEEE 12th international Conference on Nano/Molecular Medicine, Engineering (NANOMED) (pp. 146-149). IEEE.

[15] Celli, F. Cumbo, F. & Weitschek, E. (2018). are the Classification of the large DNA methylation of the datasets and also the identifying of cancer drivers

[16] Vistro, D. M., Rehman, A. U., Abid, A., Farooq, M. S., & Idrees, M. (2020). ANALYSIS OF CLOUD COMPUTING BASED BLOCKCHAIN ISSUES AND CHALLENGES. Journal of Critical Reviews, 7(10), 1482-1492.

[17] Ayer Deposition Technique To Enhance Bio-Medical Device, Hoxha, S., Shepard, A., Troutman, S., Diao, H., Doherty, J. R., Janiszewska, M., ... & Kissil, J. L. (2020). YAP-mediated recruitment of YY1 , EZH2 represses transcription of key cell cycle regulated Cancer Research.

[18] Vistro, D. M., Rehman, A. U., Farooq, M. S., Abid, A., & Idrees, M. (2020). A LITERATURE REVIEW ON SECURITY ISSUES IN CLOUD COMPUTING: OPPORTUNITIES AND CHALLENGES. Journal of Critical Reviews, 7(10), 1446-1455

[19] Image Compression Technique using Particle Swarm Optimization. on 2018 international Conference on Advanced Computation, Telecommunication (ICACAT) (pp. 1-4). IEEE.

[20] Hamedan, A. A., Boubou, E., Jamal, K., & Al-Omari, A. (2018, November). Availability, Usability of then Hospital-based Cancer Registry Data and also Measuring then Quality Outcome indicated

[21] Panted, I., & Aramco, A. (2019, October). Bio-sensed data warehouse with analytics e-health solutions. on 2019 10th IEEE international Conference on Cognitive communications (CogonfoCom) (pp. 169-174). IEEE.