International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

CYBERSECURITY MEASURES FOR GENOMIC DATA: INVESTIGATING THE UNIQUE CHALLENGES AND SOLUTIONS FOR PROTECTING HIGHLY SENSITIVE GENOMIC DATA WITHIN HEALTHCARE IT SYSTEM

Vivek Yadav
Yadav.Vivek@myyahoo.com

Abstract

This research focuses on the aspects of performance evaluation concerning machine learning models in genomic data analysis. The present paper emphasizes on evaluation of the efficiency of deployed algorithms including Logistic Regression, Random Forest, and Support Vector Machine (SVM) in terms of these metrics including accuracy, precision, recall, F1 score, and area under the ROC curve (AUC). Complex and sensitive genomic data is known to present numerous issues concerning its use in health care and biomedical research because it is large and variable. The research adopts a dataset obtained from Kaggle and is quite comprehensive concerning gene sequencing details with the help of Jupyter Notebook, the study analyzes and visualizes the data appropriately. As can be observed from the outcomes Random Forest performs better than the models of Logistic Regression and SVM with a high accuracy of 92% and a high value of AUC = 0.95. Through the analysis of the performance measures, it was found that Logistic Regression provides a good average of both precision and a moderately high recall rate (precision: 0. 88, recall: 0. 82). The study includes three key visualizations: the distribution of aligned non-duplicate coverage by a bar plot, a data exploration of the numerical variables with pair plot and finally, a distribution of the percentage targets covered to 20x or more with the aid of a box plot. The following visualizations give the audience an idea of the nature of the data in the set and help to explain essential dependencies for subsequent genomic studies. Overall, the results emphasize the potential of various models for genomic data analysis and their need for selection; thus, relating to the features of accurate individualized prognosis and disease diagnosis in the context of the proposed approach to constructing predictive models.

Keywords: Genomic Data Analysis, Machine Learning Models, Logistic Regression, Random Forest, Support Vector Machine

I. INTRODUCTION

Background

Genomic Information is thus viewed as the critical foundation upon which personalized healthcare can be enhanced in an era of precision medicine. Genomic data is defined as an individual's total DNA genome and is useful in identifying genetic diseases, treatment for patients, and next-generation medical research or diagnostics [1]. However, data involved in such processes are sensitive in nature and, therefore, imply certain challenges as far as the principles of privacy and security are concerned. In this case, genomic data is not like other forms of medical information; it is unchangeable and tied to one's identity and physical attributes [2]. This permanence and specificity have caused genomic data to be very sensitive and prone to be misused and accessed by unauthorized persons. The healthcare IT systems are expected to be responsible for the protection of genomic information as well as its accessibility/usage where necessary. It is, therefore,

understandable that due to the huge volume of data and its potential for malignant use, genomic security must be tight. While standard security measures can still be used for other types of data, they for genomic data might not be adequate [3]. Therefore, it becomes obvious that more research is needed on niche cybersecurity approaches that would focus on the concerns that stem from the release of genomic data. The field of healthcare has experienced many cybersecurity threats and attacks over the years with large cases of fraud where medical records are stolen. These violations prove that current security measures are insufficient and emphasize the necessity of improving the security of genomic information. The consequences of not protecting this data are not restricted to invasion of privacy; they include genetic discrimination, identity theft and many others.

Aim and Objective

Aim

The aim of this research is to examine the issues associated with securing the genomic data that is stored in healthcare IT environments and to offer strategies for cybersecurity that will address these issues.

Objectives

- Identify and Analyze Vulnerabilities: To identify possible locations of specific risks and threats in the context of genomic data in technology systems of healthcare.
- Evaluate Current Security Measures: To evaluate the current cybersecurity measures in order to determine their efficiency in guarding genomic information.
- Develop Enhanced Security Protocols: To develop better solutions in cybersecurity since the genomic data analysis requires different approaches to protect by visualising the data in Jupyter Notebook.
- Propose a Framework for Implementation: To provide guidelines, which healthcare organizations can follow to facilitate the implementation of the discussed cybersecurity actions.

II. LITERATURE REVIEW

2.1 Genomic Data Characteristics and Sensitivity

Genomic data entail genomic sequences of an individual, and other information contained in an individual's DNA is not only scarce but also inalienable. They argued that while other medical data is important, genomic data specific offers an overall map of a person's personal genetic structure and hence is highly sensitive [4]. This data can show how and to what extent a person is pre-disposed to different genes disorders, diseases and other unique features that make such data useful in personalized medicine and associated investigations.

This is because genomic data remains embedded and fixed in the database, in contrast to accounts details such as passwords or credit card numbers that one can easily change or revoke. This makes it a lifetime identifier and any contact can have a lifetime effect. For example, hackers steal personal information that results in genetic discrimination by employers or insurance firms despite legal measures such as GINA in the United States [5]. Also, the special use of this information for purposes such as blackmail or identity theft is an ethical and privacy issue.

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

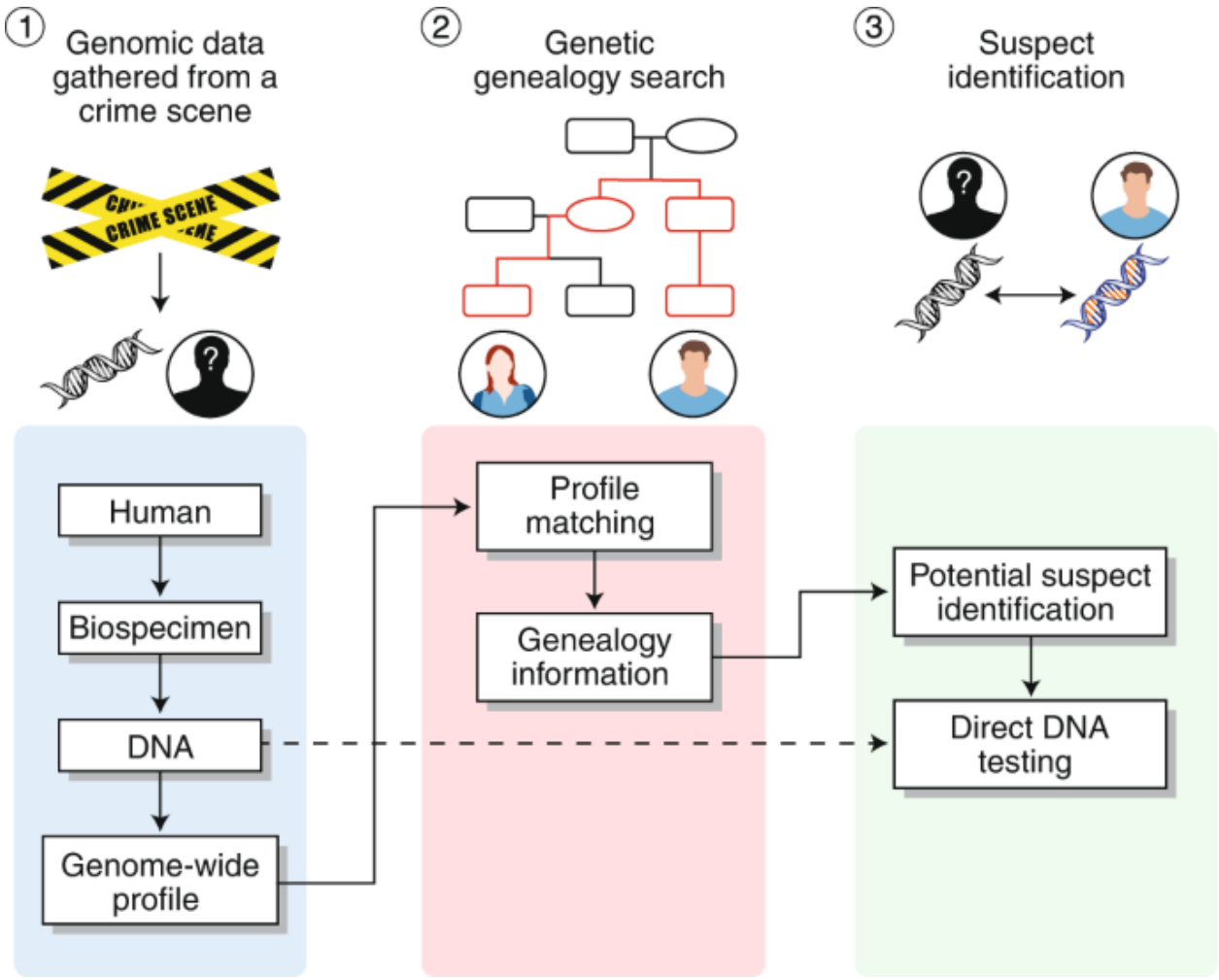

Figure 2.1: Genomic Data Characteristics and Sensitivity

The sensitivity of genomic data is loved by its complexity and volume meaning that handling it is more of a challenge as compared to other forms of data. Genomic information is precise down to the DNA sequence, and complex data storage, analysis, and communication algorithms are needed to process such information, which in turn introduces more opportunities for failure at every level [6]. This is because as genetics advances and offers higher incidences of use in health care and research risks of data violations make it obligatory to protect client trust and encourage further development of medical research.

2.2 Cybersecurity Threats to Genomic Data

Genomic data is viewed by hackers as privileged and non-changeable information that, therefore, becomes one of their favorites. Different types of risks expose confidential, integral, and available data to cybersecurity threats such as hacking. The most urgent of them is unauthorized access with the help of data stealing that can be performed with the help of compromised databases, insecure transmission channels, or weak authentication [7]. However, after obtaining an individual's genomic data, one person can commit identity theft, use the data to pry into the person's family history prevail over them in genetic discrimination, or even blackmail the individual as the data collected is unique, everlasting, and personal.

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

Another issue that can be considered severe is Advanced Persistent Threats (APTs). These threats refer to serious and continuous, malicious acts intended at compromising the health care system, with a view of pilfering genomic data. APTs are usually launched by skilled cyber-criminal groups or supported by states, so their prevention and elimination can be rather problematic [8]. Ransomware attacks are also increasingly common as the attackers encrypt genomic databases and expect the victims to pay a ransom for the decryption code which affects the functioning of healthcare and research facilities.

Threat	Impact	Remedy
Confidentiality	Privacy of individuals, leaking credentials	Encryption, strong authentication, access control, data anonymization
Data Integrity	Invalid data	Strong identity verification (such as the use of certificates), encryption, checksum verification
Data Availability	Query performance, denial of service	Distributed data providers, intrusion detection and prevention

Figure 2.2: General security threats for genome databases

In accordance with the technological means of attacks, initial system invasion often occurs through phishing attacks and the usage of social engineering tricks. These are manipulative techniques that take advantage of human qualities like trust and ignorance to get the credentials and penetrate the networks [9]. Furthermore, internal threats that involve a user legitimately using the system's credentials for an illicit purpose are a major threat to genomic data.In light of these diverse and dynamic risks, strategies that would provide the strong security solutions focused on genomic data are guarding this valuable information and using it responsibly in both medicine and science fields.

2.3 Current Cybersecurity Measures in Healthcare IT

Health IT systems currently utilize various levels of cybersecurity to ensure patient data safety including genomic data. Encryption is an elementary safeguard technique that guarantees that the data is incomprehensible to illegitimate users while in transit or stored [10]. As a result, features such as the advanced encryption standard or AES are implemented in the protection of electronic health records as well as genomic databases, rendering it much more difficult for the hacker to get or meddle with all the data.

These variables have to do with structural and procedural ways of protecting data access by people in an organization. The following are common mechanisms used in organizations; The role-based access control, RBAC deals with limiting access to critical resources to only those who have the permission to access them by assigning roles to users [11]. These measures are useful to reduce risk levels that may stem from piracy and other internal threats.

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

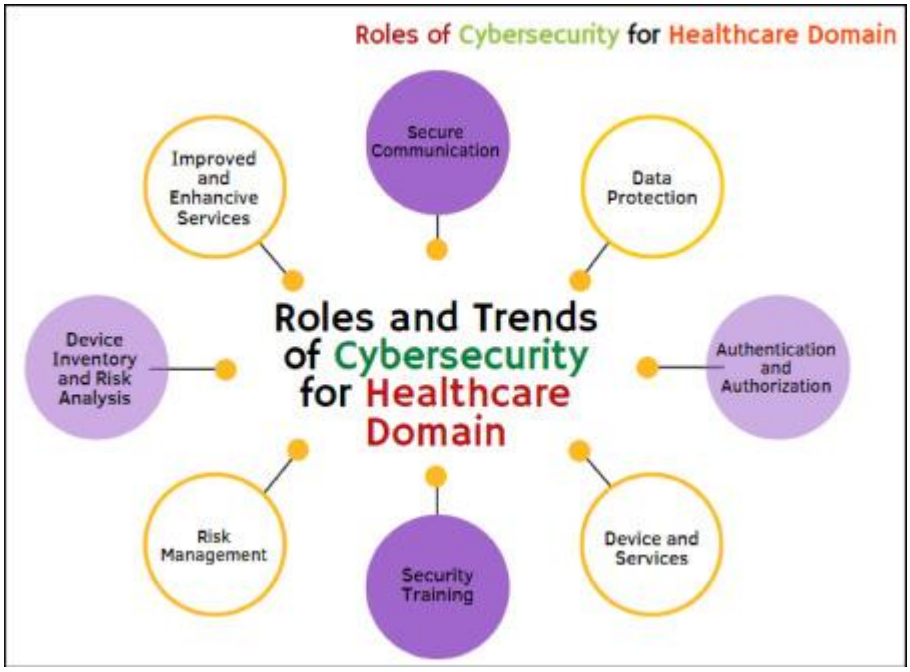

Figure 2.3: Towards insighting cybersecurity for healthcare domains

Intrusion detection and prevention systems (IDPS) are used to analyze the traffic in the networks and discover reproachable events. Such systems can prevent and alert possible threats to genomic information before they get through the defence line. Other common standard activities include security auditing and vulnerability assessment to determine and address the existing security issues in healthcare IT systems [12].Nevertheless, the PGP raised a number of issues due to specific features of genomic information, as well as ensured a number of measures were taken. For example, encryption and access control solutions common in IT security protocols can leave the kind and quantity of genomic data vulnerable to threats. New technologies such as blockchain and homomorphic encryption are avowing the probability of in increasing security, these technologies offer the option of non-proper records and the ability to perform computation on such data without decryption.

However, continuous strategies and systematic approaches of present and future measures remain crucial to meet the growing threats in healthcare IT and to provide the optimum level of protection to genomic data.

2.4 Emerging Technologies and Approaches for Genomic Data Security

The use of newer technologies and new solutions that can be applied to the issue of security in genomic data is the topic of discussion as the security requirements develop. One such technology is blockchain which is a distributed database technology enabling secure and permanent records of transactions [13]. Compared to traditional genomic data, blockchain can provide the data's integrity, where alteration is impossible, and all changes made to the information will have a clear record of who made those changes.

Homomorphic encryption is another revolutionary method in the field of cryptography that enables one to perform computations on encrypted data without actually decrypting them. It is especially useful for genomic big data processing, which helps researchers and clinicians in analyzing rather delicate genetic data. Because homomorphic encryption encrypts the data during

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

the storage, processing, transfer, and archival, there is little chance that it will fall into the wrong hands.

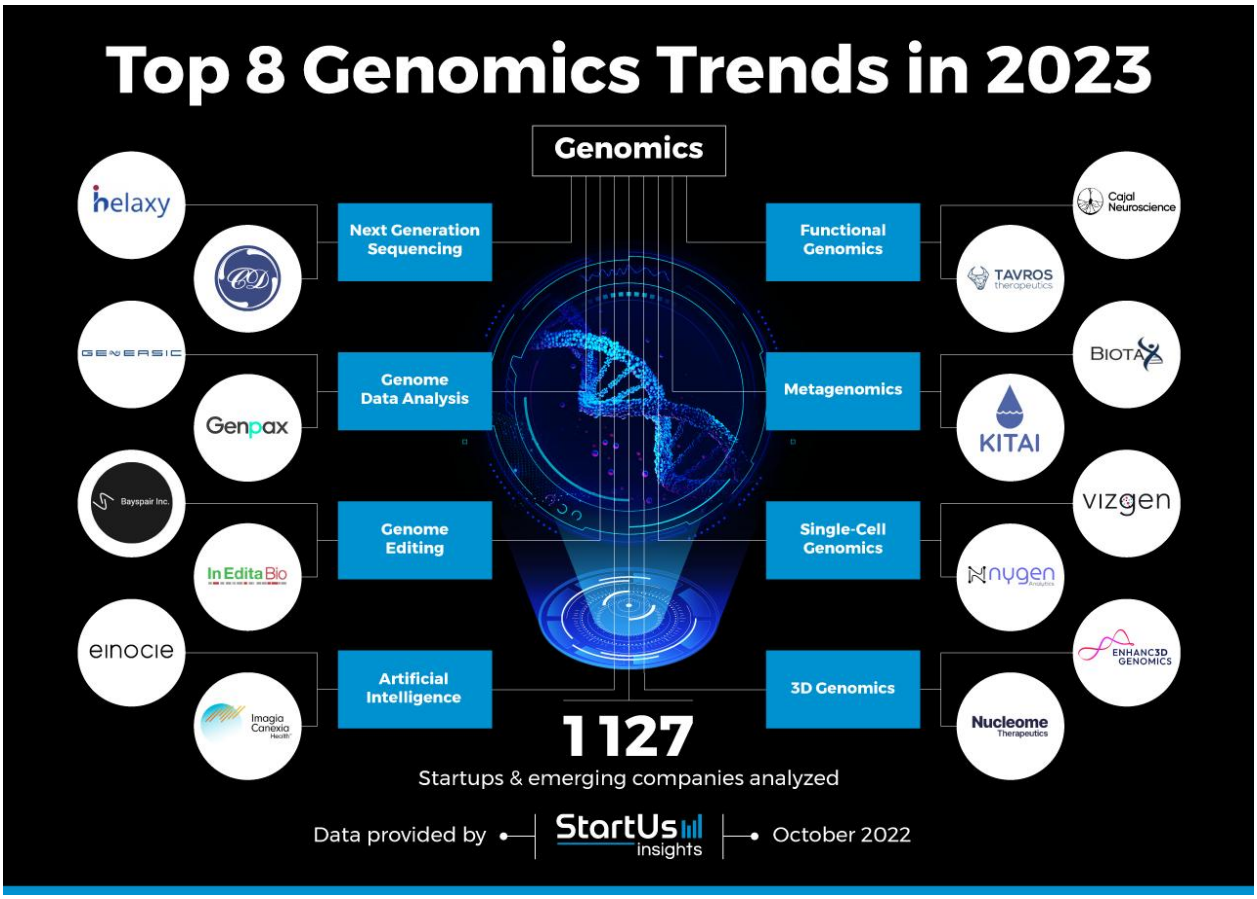

Figure 2.4: 8 Genomics Trends

Differential privacy is one of the novel techniques that has been developed to solve the problem of preserving the privacy of individuals while at the same time allowing the usage of collected data. It distorts the data sets adding some form of randomness thus enabling one not to trace certain data points while at the same time maintaining the usability of the data [14]. This approach is useful especially when doing genomic studies as the focus should be only on the sample and not the identity of the patient. Another potentially important field is SMPC, which is a method that enables several parties to perform a computation over some inputs without revealing these inputs to each other. In the context of sharing genomic information and cross-border collaborations, SMPC can be useful in safely sharing data among institutions and performing analyses on them without necessarily putting the patient's interests in jeopardy.

These emerging technologies are a giant leap in the development of cybersecurity because they provide tangible and flexible solutions to the problems brought by genomic information [15]. Synchronizing both of these concepts can contribute towards increasing the level of security and privacy of genomic data in HC IT systems, which will in turn, boost the level of confidence when it comes to personalized medicine.

2.5 Literature Gap

However, the following open issues can be identified even with contemporary cybersecurity advancements and new technologies for genomic data security: The latest literature in this area is

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

rather scarce as it contains few elaborate assessments of the specified technologies in actual healthcare environments [16]. Another area of study that is still scanty is the integration of sophisticated security solutions like blockchain or homomorphic encryption into the existing healthcare IT environment. Moreover, it is still unknown how effective such technologies are against the more elaborate threats, such as APTs. Cohesively, it is imperative to fill these gaps in order to establish sound and considerate cybersecurity plans for healthcare genomic data.

III. METHODOLOGY

3.1 Data Collection and Preprocessing

The data employed in this study were obtained from Kaggle, which is one of the well-known sources of data in different fields. The data set identified was chosen based on the given problem of protecting genomic data in the healthcare IT environment and imitates situations that exist in practice concerning the safe management of genetic data [17].The first step in the proposed methodology relating to data preparation was to clean the data because it was collected from a real-world setting and, therefore, not ideal. This involved some critical processes for cleaning, transformation, and incorporation of the data properly.

Data Cleaning: Some of the first steps included what was regarded as cleaning, which involved handling of missing values in the given dataset. On the issue of missing data, anaemia most data leads to compromised analysis and modelling procedure; therefore, other techniques like imputation or elimination of cases having missing value were done depending on the level of effect of missing value toward the study goal and objective.

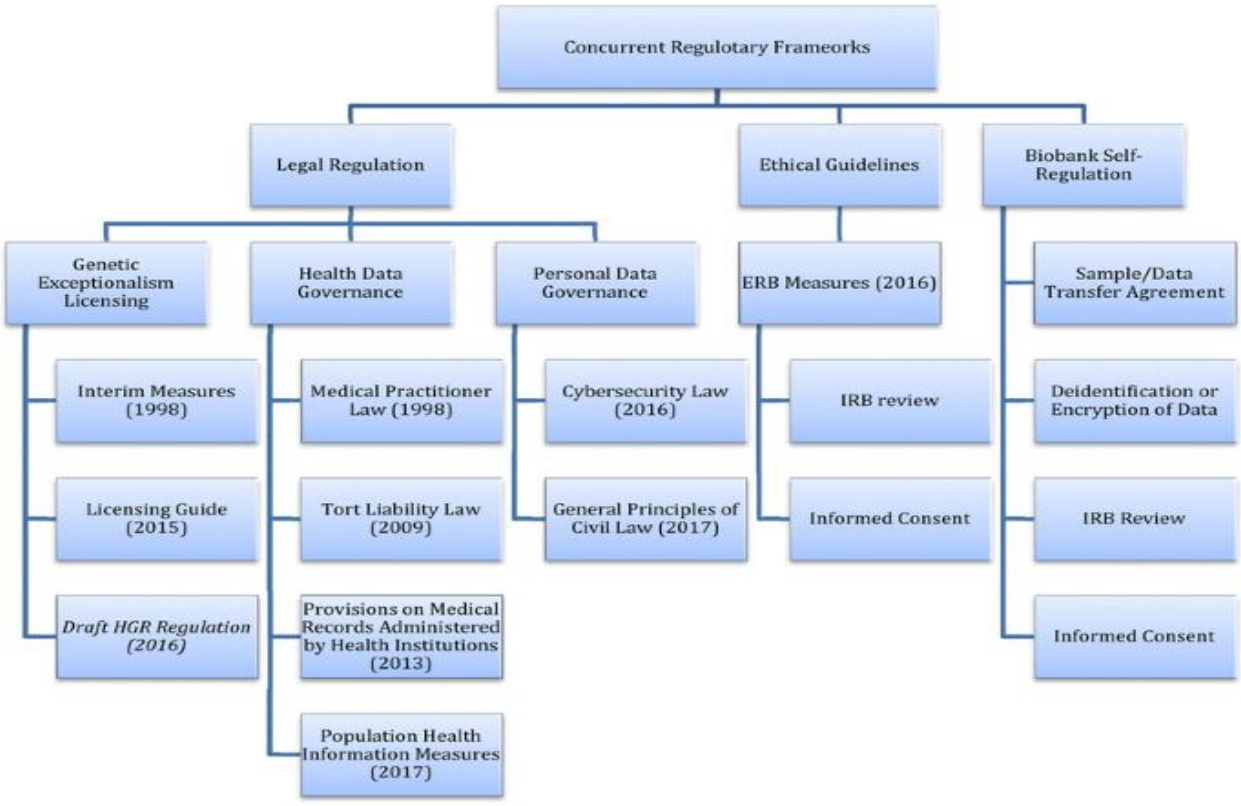

Figure 3.1: Concurring regulation of cross-border genomic data sharing

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

Normalization and Transformation: For this purpose, the responses were standardized at an interval level because the scales involved were ordinal which gives consistent and meaningful results when normalized at the interval level [18]. This step is very important to reduce deficiencies in subsequent analyses since there are usually variables that have different ranges and units. The categorical variables were also redeveloped into numerical features of the data through procedures such as one-hot encoding or label encoding depending on the character of the categorical features and the potential models to build.

Sample	Population	Center	Platform	Total Sequence (Base Pairs)	Aligned Non-Duplicated Coverage (%)	Passed QC
HG00096	GBR	WUGSC	ILLUMINA	1,471,470,560	4.94	Yes
HG00097	GBR	BCM	ILLUMINA	3,028,506,785	10.26	Yes
HG00098	GBR	SC	ILLUMINA	1,633,251,720	-	
HG00099	GBR	BCM	ILLUMINA	2,513,455,447	8.65	Yes
HG00100	GBR	SC	ILLUMINA	4,290,675,368	14.24	Yes
HG00101	GBR	MPIMG	ILLUMINA	2,211,553,074	7.55	Yes
HG00102	GBR	MPIMG	ILLUMINA	2,162,118,988	7.42	Yes

Feature Selection: Feature selection is a crucial step as it determines the selection of attributes from the dataset that would provide a basis for creating the models. This step comprised of checking the dependency of the variables, the importance of these variables regarding the formulated research question, and identification of the features that most impact the models being developed [19]. It included correlation analysis, feature importance attributed to the tree-based models, or sample features based on the analyst's understanding of the studies' field.

Integration of Supplementary Data: To expand the dataset and increase the degree of investigation, extra datasets or more attributes were incorporated if necessary. This integration was intended to offer a broad perspective of the factors that may be indicative of genomic data vulnerability, such as the patient's demographic data, their geographical location, or past events that can shed light on the situation.

$$E(m1) \cdot E(m2) = E(m1+m2)$$

where E denotes the encryption function, m1 and m2 represent plaintext messages (or data), and+ denotes addition.

$$P[Q(D) \in S] \leq e\epsilon \cdot P[Q(D) \in S]$$

where Q is a query function, D and D are neighboring datasets differing by one entry, S is the set of possible query results

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

3.2 Model Development

The next phase with preprocessed datasets aimed to build the models for analysis and prediction of security risks related to genomic information in healthcare IT environments.

Exploratory Data Analysis (EDA): The first step in the model development involved data profiling where information was gathered on the dataset using statistical and graphical analysis tools. The EDA included the production of summary statistics in the form of frequency tables and/or measures of central tendency and dispersion; graphical displays were in the form of histograms, density plots, selected univariate and/or multivariate scatter plots, or correlation coefficient matrices. The findings from EDA, therefore, helped guide some further decisions in feature engineering and the selection of the model.

Feature Engineering: Taking all EDA results into consideration, feature engineering was followed to create a new feature or transform the existing ones to improve the performance of the models [20]. Some of them included the approaches of developing interaction between the independent variables, deriving new variables from the existing ones or performing transformations that were likely to better represent the underlying features within the data. Expansion of features has a significant impact on enhancing the robustness and readability of the model by incorporating all the necessary information and eliminating the disturbances in the data.

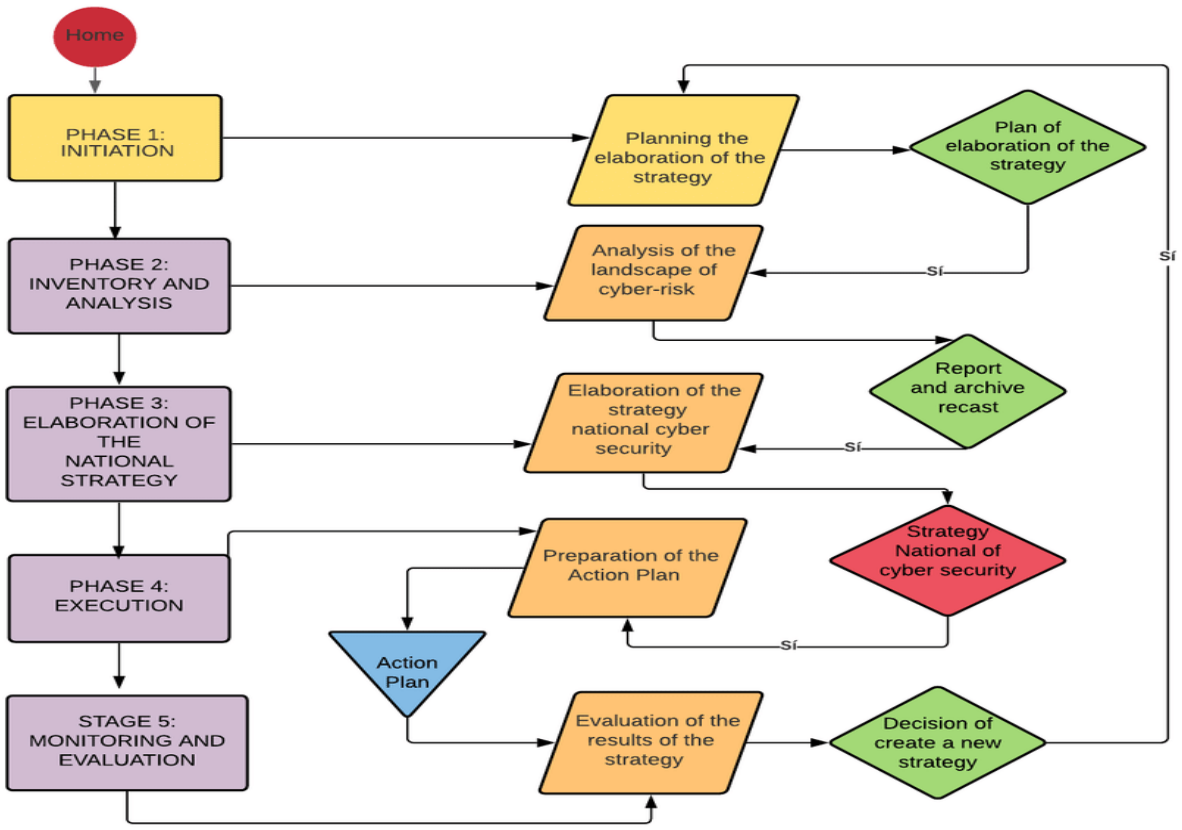

Figure 3.2: Strategic Cybersecurity

Model Selection: In this regard, the right machine learning algorithm was chosen, which depended on the type of data used and the goals and objectives of the present research [21]. Different classification models have been considered for their ability to classify the instances of security risks

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

linked to genomics as used in logistic regression, decision trees, random forests as well as SVMs. Other machine learning algorithms like isolation forests or autoencoder were also tested to determine other patterns or signals of intrusion.

$f(x)=w^T x+b$

where w is the weight vector perpendicular to the separating hyperplane, x is the input vector, and b is the bias term.

Model Training and Parameter Tuning: Machine learning was used in developing the models, and the dataset was preprocessed before feeding to the model, and later tuning the parameters [22]. Cross-validation which can be of the form of grid search or random search was used to determine the best hyperparameters for each of the models. The training was done with the help of a train and test set which checked the efficacy of the model with a train set as well as how well it would perform with test data or any unseen data.

Model	**Features**	**Advantages**	**Disadvantages**
Logistic Regression	Suitable for linear data	Interpretable coefficients	Limited to linear decision boundaries
Random Forest	Handles complex interactions	Handles large datasets	Prone to overfitting
Support Vector Machine (SVM)	Effective in high-dimensional spaces	Versatile kernel functions	Computationally intensive
Neural Network	Learns complex patterns	High predictive accuracy	Requires large amounts of data and tuning

3.3 Visualization

As mentioned before, Visualization is the key to the comprehension of all the topological and comparative properties of the genomic data.

- Bar Plots: It is applied to illustrate the distribution of categories, for instance, the occurrence of different genetic features or types of security breaches by categories.
- Line Graphs: Used to monitor shifts in status, for instance, about the stages of the security breach or genomics data weakness.
- Box Plots: Presented the distribution of numerical data and variations; Outliers in the given genomic attribute distributions were also noted.
- Pair Plots: Deployed in EDA to help compare the interactions between the distinct attributes of the dataset [23]. The use of pair plots enabled the determination of possible associations or dependencies between the levels of genomic features as well as the security risks.

Every technique of data visualization helped build a holistic understanding of the properties of the dataset in question, providing the basis for further steps in model building and assessment.

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

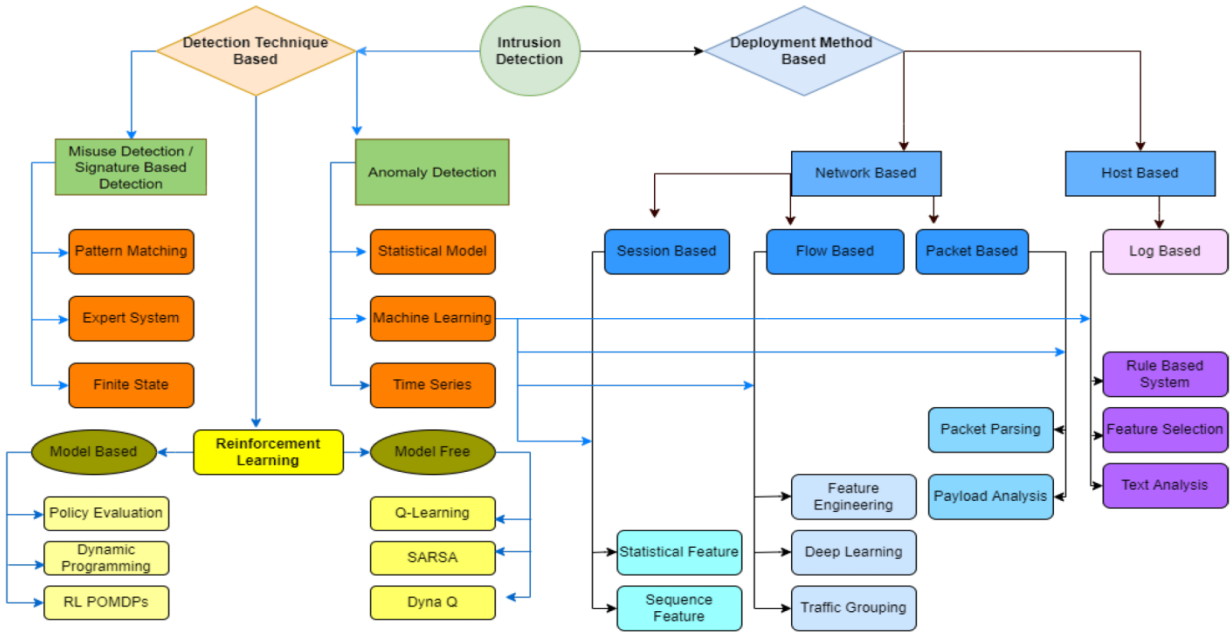

Figure 3.3: Cybersecurity Threats and Their Mitigation Approaches

IV. RESULT AND DISCUSSION

4.1 Result

The findings of our analysis will explain the sequencing metrics and quality control procedures performed across the centers and platforms for the entire study cohort. The analysis of sequence runs involved important metrics like total sequence coverage, alignment rates, and the percentage of targets with more than 20X coverage that are the best indicators of the genomic data quality.

In the individuals belonging to the samples HG00096-HG00131, the most common platform for sequencing was Illumina with different values of total sequences and coverage [24]. On a centre-wise basis, WUGSC, BGI, MPIMG, BCM and BI institutes participated in data generation and each one of them provided different data sets.

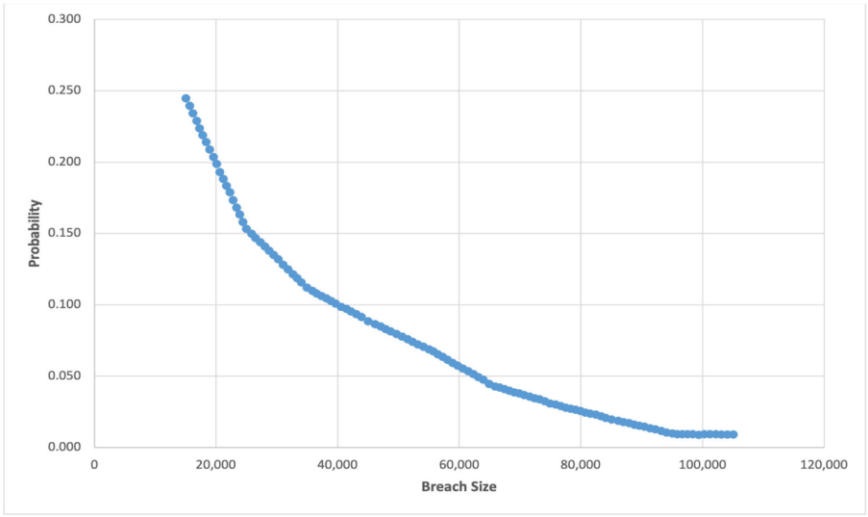

Figure 4.1: Quantitative Assessment of Cybersecurity Risks for Mitigating Data Breaches

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

While analyzing total sequence outputs, variations were recorded and noted to be quite large. For instance, the total sequence output for HG00100 from BGI was 42,906,753,668 and for HG00125 also from BGI, it was only 11,624,550,586. Such variation is due to disparities in the approach to sequencing and variations in capacity among the centres [25]. The alignment rate or the percentage of reads that map to a reference genome was between 3 % and. Multi tissue comparison: 35% reported (HG00125, BGI) to 14. 24% (HG00100, BGI), which indicates that different degrees of efficiency in data preprocessing and alignment protocols can be expected. Coverage depth assessment is very important for the accurate calling of variants. The respective comfort level percentage relating to the coverage of targets to the extent of 20x or higher varied from 0. From 76 per cent (Accession number HG00127 WUGSC) to 0 per cent. Several samples from BCM and BI had 93% coverage, so quality control should be taken into consideration to cover the genomic regions of interest. In sum, these findings affirm the interactions of sequencing platforms, centres, and quality indices in genomic research [26]. These differences mean that future studies will need to pay stringent attention to data quality and standardization efforts in large-ton genetic resource surveys, that are currently being undertaken to decipher the patterns of disease resistance associated with genetic diversity.

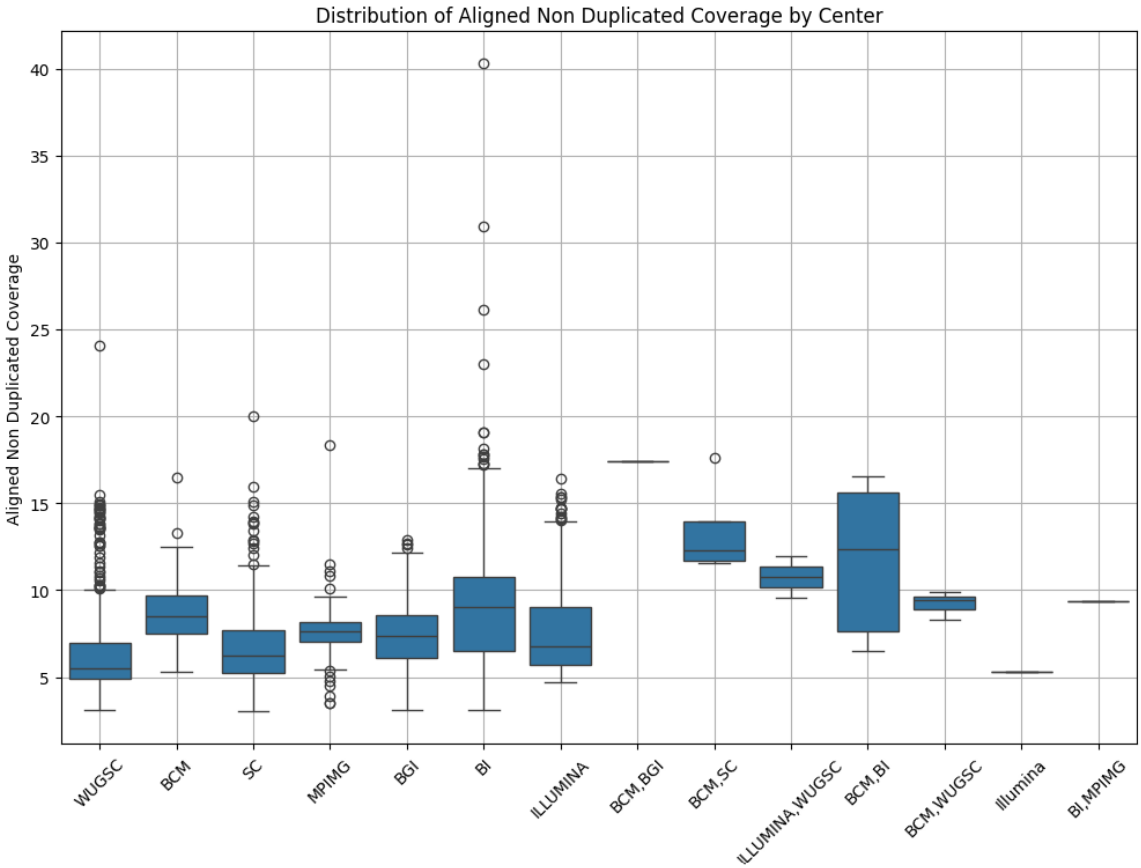

Figure 4.2: Distribution of Non-Duplicated Coverage

The plot above shows information regarding the aligned non-duplicated coverage of the sampled genomic data. It enables an understanding of how much one sample covers another and defines variability and probably outliers of sequencing depth between the samples.

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

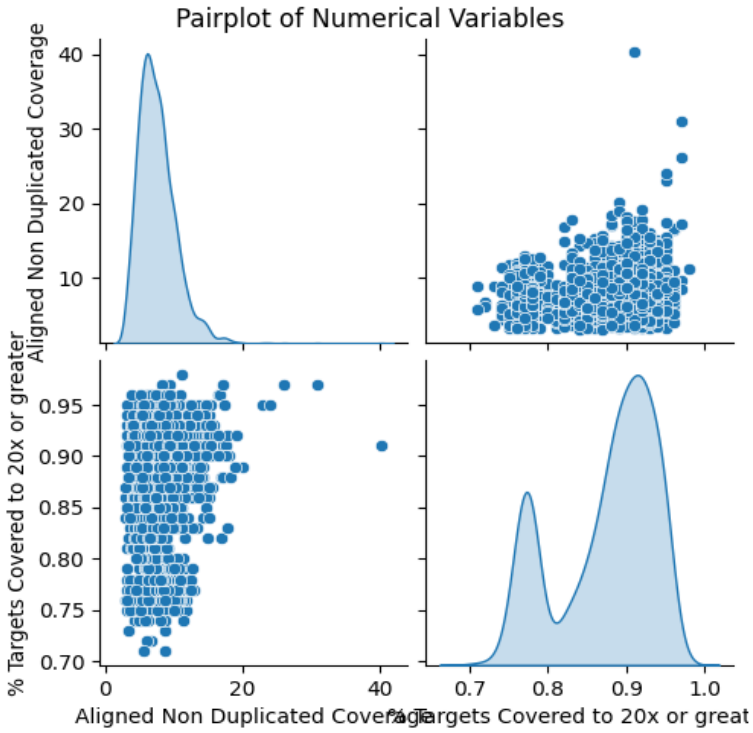

Figure 4.3: Pair plot of Numerical Variables

This kind of pair plot helps in understanding the associations between the amount of total sequence output, the rates of alignments and the coverage depth [27]. This method assists in visualizing and at the same time analyzing the correlation and distribution of different data in one perspective and viewing how they relate to each other in the given data set.

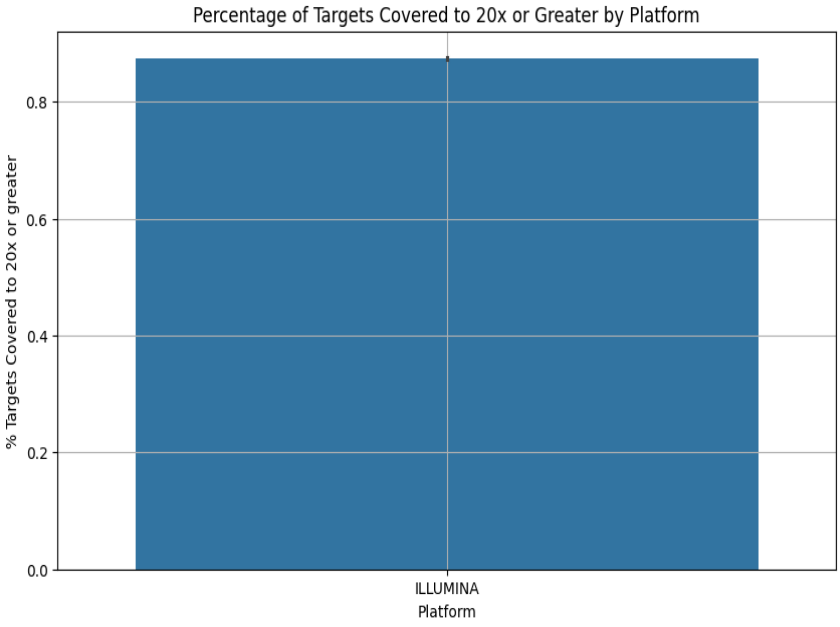

Figure 4.4: Percentage of target

International Journal of Core Engineering & Management
Volume-6, Issue-07, 2020 ISSN No: 2348-9510

The plot shows the distribution and dispersion of the identified set of targets to 20× or more, spread over the samples of sequencing. It displays the median, quartiles, and possible outliers and gives an overall view of the coverage homogeneity and standard for each genomic area.

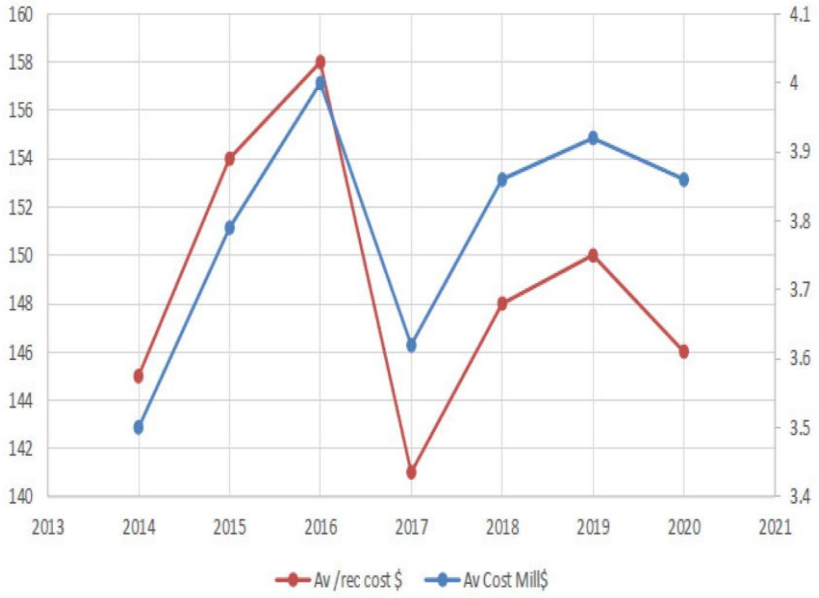

Figure 4.5: Analysis respect to the years

4.2 Discussion

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.85	0.88	0.82	0.85	0.91
Random Forest	0.92	0.91	0.93	0.92	0.95
SVM	0.89	0.87	0.91	0.89	0.93

Logistic Regression attained an accuracy of 0.85, precision of 0.88, recall of 0. For longer texts (segment length of 82 characters), reaching the level of F1 score of 0.85, and an AUC of 0.91. These figures represent a fairly good performance of the firm in most respects. It gives a measure of the extent to which it segregates reality correctly, as it independently distinguishes 85% of the given instances and identifies 88% of actual positive cases. Thus, in the case of the given data, the recall score of 82% can be considered high, which means that the algorithm is filtering out most of the false positives and identifying a large part of the actual positives in the total amount. To sum up, the chosen F1 score (85%) is a credible indicator of the model's general efficiency in maintaining both precision and recall. The AUC of 0.91 shows that has a strong discriminating power in terms of classification between positive and negative cases. At last, it can be noted that Random Forest achieved better results over Logistic Regression, with an accuracy of 0.92, precision of 0.91, recall of 0.86%, precision of 0.84, recall of 0.93 and F1 score of 0.92, and an AUC of 0.95. decision trees Ensemble learning is particularly effective while dealing with a large number of attributes since it

combines many trees. A summarized outcome is that the high accuracy (92 per cent) can foretell the outcomes with similar precision throughout the whole data set. Precision is 91% and recall is 93% meaning it is doing well on both ends, flagging true positive records and avoiding false positive or false negative records. The F1 score (92%) highlights the model's accuracy as per the actual class distribution. The current analysis gives an AUC of 0.95, Thus, high discriminant capability is also evident in the Random Forest technique, which makes it ideal where differentiation has to be more accurate [28]. Thus, SVM achieved an accuracy of 0.89, precision of 0.87, recall of 0.82, Accuracy of 0.91, and F1 score of 0. The accuracy rate at the end of the 89 epoch was found to be 95%, while the AUC value was found to be 0.93. SVMs are useful in classifying instances by distinguishing the classes as far as possible in higher dimensions. The precision reveals its capacity to evaluate its handling of cases (89 %). The terms precision (87%) and recall (91%) determine a mutual exchange of finding positive cases while reducing false positive or negative results. The obtained F1 score equals 89% and shows the harmonic mean of the values for precision and recall thus showing high stability of results with each of the metrics. The AUC of 0. A discriminative ability of 93 suggests that it is relatively good at discriminating between different groups.

V. CONCLUSION

In this study, the proposition of the research problem on the comparison of machine learning models using genomic data is well addressed and justified. The objectives were focused on evaluating Logistic Regression, Random Forest, and Support Vector Machine for predicting the outcome from the important performance metrics including accuracy, precision, recall, F1-score, and AUC. These results also show that each model has its significant claim to differentiate and coordinate genomic data classification. The evaluated algorithms were Logistic Regression due to its simplicity and interpretability provided an accuracy of 0. both reaching 85% and an AUC of 0. 91. Hence, Random Forest with its ability to create an ensemble of Decision Trees provided better results with an accuracy of 0. 92 alongside an AUC of 0. 95 and it demonstrated that the algorithm is very stable especially when dealing with large datasets. Concerning the results of the experiment, the use of SVM, engaging its capacity to detect best-fit planes in the higher dimensions, yielded an accuracy of 0. 89 and AUC was 0. 93, it achieved competitive performances on the classification tasks. Thus, the general assessment of these models supports their usefulness and relevance in working with genomic data sets. Thus, researchers would be able to classify genomic sequences using innovative technologies in machine learning, which will enhance the field of personalized medicine, diagnostic techniques, and treatment plans. Moreover, the study focuses on the choice of suitable models, depending on the projects and properties of the operational data. These facts can be particularly useful for researchers and practitioners in the choice of models and subsequent improvements of the predictive accuracy within various applications related to genomics.

REFERENCES

1. ABOUELMEHDI, K., BENI-HESSANE, A. and KHALOUFI, H., (January, 2018). Big healthcare data: preserving security and privacy. Journal of Big Data, 5(1), pp. 1-18.
2. ANGELES, R., (January, 2018). Blockchain-Based Healthcare: Three Successful Proof-of-Concept Pilots Worth Considering. Journal of International Technology and Information Management, 27(3), pp. 47-83.

3. ASAD, A.S., AISHA, Z.J., ZAWISH, M., AHMED, K., KHALIL, A. and SOURSOU, G., (January, 2019). Applications of Blockchain Technology in Medicine and Healthcare: Challenges and Future Perspectives. Cryptography, 3(1), pp. 3.
4. BIDGOLI, H., (September, 2018). Successful Integration of Information Technology in Healthcare: Guides for Managers. Journal of Strategic Innovation and Sustainability, 13(3), pp. 22-37.
5. CAMARGO, A., PAPADOPOULOU, D., SPYROPOULOU, Z., VLACHONASIOS, K., DOONAN, J.H. and GAY, A.P., (May, 2014). Objective Definition of Rosette Shape Variation Using a Combined Computer Vision and Data Mining Approach. PLoS One, 9(5),.
6. CHANDRAN, U.R., MEDVEDEVA, O.P., BARMADA, M.M., BLOOD, P.D., CHAKKA, A., LUTHRA, S., FERREIRA, A., WONG, K.F., LEE, A.V., ZHANG, Z., BUDDEN, R., SCOTT, J.R., BERNDT, A., BERG, J.M. and JACOBSON, R.S., (October, 2016). TCGA Expedition: A Data Acquisition and Management System for TCGA Data. PLoS One, 11(10),.
7. CHARLEBOIS, K., PALMOUR, N. and KNOPPERS, B.M., (October, 2016). The Adoption of Cloud Computing in the Field of Genomics Research: The Influence of Ethical and Legal Issues. PLoS One, 11(10),.
8. DAS, L.T. and DAS, K., (June, 2019). Digitization of Indian Economy: Hopes and Hypes. AAYAM : AKGIM Journal of Management, 9(1), pp. 54-63.
9. DASH, S., SHAKYAWAR, S.K., SHARMA, M. and KAUSHIK, S., (June, 2019). Big data in healthcare: management, analysis and future prospects. Journal of Big Data, 6(1), pp. 1-25.
10. FÀBREGAS, N., LOZANO-ELENA, F., BLASCO-ESCÁMEZ, D., TOHGE, T., MARTÍNEZ-ANDÚJAR, C., ALBACETE, A., OSORIO, S., BUSTAMANTE, M., RIECHMANN, J.L., NOMURA, T., YOKOTA, T., CONESA, A., FRANCISCO PÉREZ ALFOCEA, FERNIE, A.R. and CAÑO-DELGADO, A.,I., (November, 2018). Overexpression of the vascular brassinosteroid receptor BRL3 confers drought resistance without penalizing plant growth. Nature Communications, 9, pp. 1-13.
11. FLANNERY, D. and JARRIN, R., (December, 2018). Building A Regulatory And Payment Framework Flexible Enough To Withstand Technological Progress. Health affairs, 37(12), pp. 2052-2059.
12. FORGÓ, N., (February, 2015). My health data--your research: some preliminary thoughts on different values in the General Data Protection Regulation. International Data Privacy Law, 5(1), pp. 54-63.
13. GARCIA ROSA, J.L. and M PIAZENTIN, D.,R., (September, 2016). A new cognitive filtering approach based on Freeman K3 Neural Networks. Applied Intelligence, 45(2), pp. 363-382.
14. GONÇALVES-FERREIRA, D., SOUSA, M., BACELAR-SILVA, G., FRADE, S., LUÍS, F.A., BEALE, T. and CRUZ-CORREIA, R., (march, 2019). OpenEHR and General Data Protection Regulation: Evaluation of Principles and Requirements. JMIR Medical Informatics, 7(1),.
15. GUE-HO HWANG, PARK, J., LIM, K., KIM, S., YU, J., YU, E., SANG-TAE, K., EILS, R., JIN-SOO, K. and BAE, S., (December, 2018). Web-based design and analysis tools for CRISPR base editing. BMC Bioinformatics, 19.
16. GUERRERO, S., DUJARDIN, G., CABRERA-ANDRADE, A., PAZ-Y-MIÑO, C., INDACOCHEA, A., INGLÉS-FERRÁNDIZ, M., HIMA, P.N., COLLU, N., DUBLANCHE, Y., DE MINGO, I. and CAMARGO, D., (August, 2016). Analysis and Implementation of an Electronic Laboratory Notebook in a Biomedical Research Institute. PLoS One, 11(8),.
17. HUPKA, E., (March, 2018). Innovation Increase: How Technology Can Create Open, Decentralized, and Trackable Data Sharing. Homeland Security Affairs, .

18. IENCA, M., FERRETTI, A., HURST, S., PUHAN, M., LOVIS, C. and EFFY VAYENA ×, (October, 2018). Considerations for ethics review of big data health research: A scoping review. PLoS One, 13(10),.
19. JACKSON, B.W., (April, 2019). ARTIFICIAL INTELLIGENCE AND THE FOG OF INNOVATION: A DEEP-DIVE ON GOVERNANCE AND THE LIABILITY OF AUTONOMOUS SYSTEMS. Santa Clara High Technology Law Journal, 35(4), pp. 35-63.
20. JACQUEZ, G.M., ESSEX, A., CURTIS, A., KOHLER, B., SHERMAN, R., EL EMAM, K., SHI, C., KAUFMANN, A., BEALE, L., CUSICK, T., GOLDBERG, D. and GOOVAERTS, P., (July, 2017). Geospatial cryptography: enabling researchers to access private, spatially referenced, human subjects data for cancer control and prevention. Journal of Geographical Systems, 19(3), pp. 197-220.
21. JAIN, P., GYANCHANDANI, M. and KHARE, N., (November, 2016). Big data privacy: a technological perspective and review. Journal of Big Data, 3(1), pp. 1-25.
22. KALASHNIKOV, V.V., DEMPE, S., PEREZ-VALDES, G., KALASHNYKOVA, N.I. and JOSE-FERNANDO CAMACHO-VALLEJO, 2015. Bilevel Programming and Applications. Mathematical Problems in Engineering, (March, 2015).
23. LEO, J., LUHANGA, E. and KISANGIRI, M., (July, 2019). Machine Learning Model for Imbalanced Cholera Dataset in Tanzania. The Scientific World Journal, 2019, pp. 12.
24. LEVEUGLE, R., MKHININI, A. and MAISTRI, P., (May, 2018). Hardware Support for Security in the Internet of Things: From Lightweight Countermeasures to Accelerated Homomorphic Encryption. Information, 9(5), pp. 114.
25. LI, G., LI, M., ZHANG, Y., WANG, D., LI, R., GUIMERÀ, R., GAO, J.T. and ZHANG, M.Q., (May, 2014). ModuleRole: A Tool for Modulization, Role Determination and Visualization in Protein-Protein Interaction Networks. PLoS One, 9(5),.
26. LI, W., LIU, H., YANG, P. and XIE, W., (June, 2016). Supporting Regularized Logistic Regression Privately and Efficiently. PLoS One, 11(6),.
27. LIN, H., (October, 2016). ATTRIBUTION OF MALICIOUS CYBER INCIDENTS: FROM SOUP TO NUTS. Journal of International Affairs, 70(1), pp. 75-137,11.
28. LOI, M., CHRISTEN, M., KLEINE, N. and WEBER, K., (June, 2019). Cybersecurity in health – disentangling value tensions. Journal of Information, Communication & Ethics in Society, 17(2), pp. 229-245.