BIGDATA ANALYTICS IN REAL TIME: SECURITY AND PRIVACY CONCERNS

[1]G.Vedavyasa, [2]A.Ragavendra Rao, [3]B.Sumalatha, [4]Gajjela Raja Shekar
[1,2,3]Assistant Professor, [4]UG Student, [1,2,3,4]Department of Computer Science Engineering, Brilliant Grammar School Educational Society Group of Institutions Integrated Campus, Hyderabad, India

ABSTRACT
Big data describes extremely large data collections with a more complex and varied structure. These characteristics typically result in more difficulties while storing, analyzing, using additional procedures, or extracting data. The practice of analyzing vast amounts of complex data to find hidden patterns or identify hidden links is known as "big data analytics." Yet the growing use of big data and its security and privacy are clearly at odds. Covers the most essential components of the organization and management of computer infrastructures to fulfill the most important security requirements of big data applications one of them is privacy. It is an important issue to solve since people are increasingly sharing personal data and content on social networks and public clouds through their smartphones and PCs. As a result, developing a safe framework for social networks is a prominent research topic. This final topic is addressed in one of the current chapter's two case study parts. Furthermore, typical security techniques such as firewalls and demilitarized zones are unsuitable for use in Big Data computing systems. However, if data is not adequately safeguarded from risks such as phishing, hacking, and so on, These challenges include weaknesses in public databases; protection against security breaches and data leaks; and so on. The security and privacy regulations create a significant problem in tracking and monitoring data access and usage in a dynamic, decentralized environment when handling large-scale, distributed data sets. The goal of this work is to look into the problems that come up when trying to keep big data secure and private, as well as to find ways to deal with these problems.
Keywords: Big Data, Security, Privacy, Data Ownership, Cloud, Social Applications, Intrusion Detection, Intrusion Prevention.

INTRODUCTION
The phrase "big data analytics" describes the use of advanced analytical techniques on datasets that range in size from a terabyte to a zettabyte, are diverse in their data, and have both a large and large variety of origins.
What exactly does the phrase "big data" refer to? Big data in this sense refers to data sets that are too huge or complicated for conventional relational databases to effectively store, manage, and analyze. The vast volume, high velocity, and great diversity of its data sets are what constitute big data. Data sources are becoming increasingly diverse and complicated as a result of artificial intelligence (AI), mobile devices, social media, and the Internet of Things, making them more challenging to handle than conventional data sources. (IoT). Sensors, devices, video/audio, networks, log files, transactional applications, the web, and social media are all sources of data, and much of it is created in real time and at a massive scale.
Better and quicker decision-making, predictive modeling, and corporate intelligence are all possible with the help of big data analytics. To analyze and store the ever-increasing amounts of data, open source technologies like Apache Hadoop and Apache Spark should be considered as part of your big data solution's architecture.

Big data poses big privacy issues.
Data owners must keep up with the rate of data growth and the plethora of rules that govern it in the era of multi-cloud computing, especially those that protect the privacy of sensitive data and personally identifiable information (PII).

The commercial risk of a privacy breach has never been bigger due to the spread of more data across more websites. Risks associated with this include hefty penalties and a decrease in market share.

When it comes to the privacy of big data, customer trust is a worry. It will be easier to "connect the dots" and understand users' past, current, and potential future behavior the more information you gather about them. You will eventually be able to compile thorough profiles of each user's preferences and way of life. Being upfront and truthful with your customers about what you're doing with their information, how you're maintaining it, and what steps you're doing to follow by privacy and data protection rules becomes more important as you gather more data. Legacy systems and e-commerce are two examples of existing sources of data that have a constantly expanding volume and velocity of data. Two other fresh (and growing) types of data sources are streams from IoT devices and social networks. In order to keep up, your big data privacy strategy must likewise grow. That requires you to consider each of the following:

How will your data security be scalable to counter insider threats and the increasing frequency and size of data breaches?

The first forecast is the expansion of regulations protecting data privacy. As more categories of sensitive data are held in larger amounts for longer periods of time, organisations will face more pressure to be honest about the data they gather, how they use and analyse it, and why they must keep it. One well-known example is the General Data Privacy Regulation (GDPR) of the European Union. Increasing numbers of governmental organisations and regulatory agencies are beginning to follow suit. Companies need reliable, scalable privacy tools for big data that support and encourage users to access, review, rectify, anonymize, or even completely erase their sensitive and personal data in order to fulfil these rising demands.

Prediction 2: Businesses will be able to delve deeper into historical data, discover applications for it that weren't initially envisaged, and combine it with new data sources with the aid of new tools for studying big data.

Using big data analytics tools and solutions, which have access to data sources that were previously unreachable, it is now possible to uncover new connections hidden in historical data. Gaining a thorough understanding of your company data is quite advantageous, especially for customer 360 and analytics activities. However, it also raises concerns about how reliable obsolete data is and how simple it is to track down organisations and get authorization to use their data creatively. You can protect your big data's privacy while simultaneously increasing its worth by regularly reviewing these four fundamental data management activities. Data collection Keeping and archiving Usage of data, including data masking options in testing, DevOps, and other applications. Setting up and upholding disclosure policies and processes Companies that have a reliable, scalable data governance programme will be in a better position to evaluate these duties because they will be able to rapidly and properly assess the risks and benefits of data-related data and make more informed decisions. preparing for the protection and privacy of big data While current multi-cloud architectures scatter data over more locations and data kinds than ever before, independent of platform, traditional data security is focused on the system and network.

Big data privacy issues cannot be disregarded. It must be a crucial element of your strategy for managing and integrating your cloud data.

In order to make it obvious which data is crucial, why it is crucial, who owns it, and how it may be used responsibly, data governance standards must be established and monitored.

You must locate, classify, and comprehend a sizable scale of sensitive data across all big data platforms using artificial intelligence and machine learning tools to automate controls. You may then utilise the rules you created for handling big data once you get this information.

Data subjects and identities need to be catalogued, listed, and connected in order to facilitate data access rights and notifications.

If you want to determine how exposed you are to risk, decide how to prioritise investments and resources for data protection, and develop plans for protection and problem-solving as your big data increases, you must be able to do continuous risk analysis for sensitive data.

To simplify and make it simpler to govern data access, such as by looking at, altering, and putting in place access controls, big data privacy tools that connect with native big data tools like Cloudera Sentry, Amazon Macie, and Hortonworks Ranger are required.

You need scalable, rapid, and effective data protection tools, such as dynamic masking for big data

used in production and data lakes, permanent masking for big data used in development and analytics settings, and encryption for big data at rest in data lakes and warehouses.

You must evaluate and explain the status of the main data privacy risk indicators as an essential part of tracking progress in protecting sensitive information and encouraging audit readiness. Click here for more information on how Informatica incorporates big data privacy into data governance and compliance.

RECENT TECHNIQUES OFPRIVACY PRESERVING INBIG DATA

Differential privacy

Differential Privacy is a technology thatprovides researchers and database analysts a facility to obtain the useful information from the databases that contain personal information of people without revealing the personal identities of the individuals. This is done by introducing a minimum distraction in the information provided by the database system. The distraction introduced is large enough so that they protect the privacy and at the same time small enough so that the information provided to analyst is still useful. Earlier some techniques have been used to protect the privacy, but proved to be unsuccessful.

In mid-90s when the Commonwealth ofMassachusetts Group Insurance Commission (GIC) released the anonymous health record of its clients for research to benefit the society . GIC hides some information like name, street address etc. so as to protect their privacy. Latanya Sweeney (then a PhD student in MIT) using the publicly available voter database and database released by GIC, successfully identified the health record by just comparing and co-relating them. Thus hiding some information cannot assures the protection of individual identity.

Differential Privacy (DP) deals to provide the solution to this problem as shown Fig. 4. In DP analyst are not provided the direct access to the database containing personal information. An intermediary piece of software is introduced between the database and the analyst to protect the privacy. This intermediary software is also called as the privacy guard.

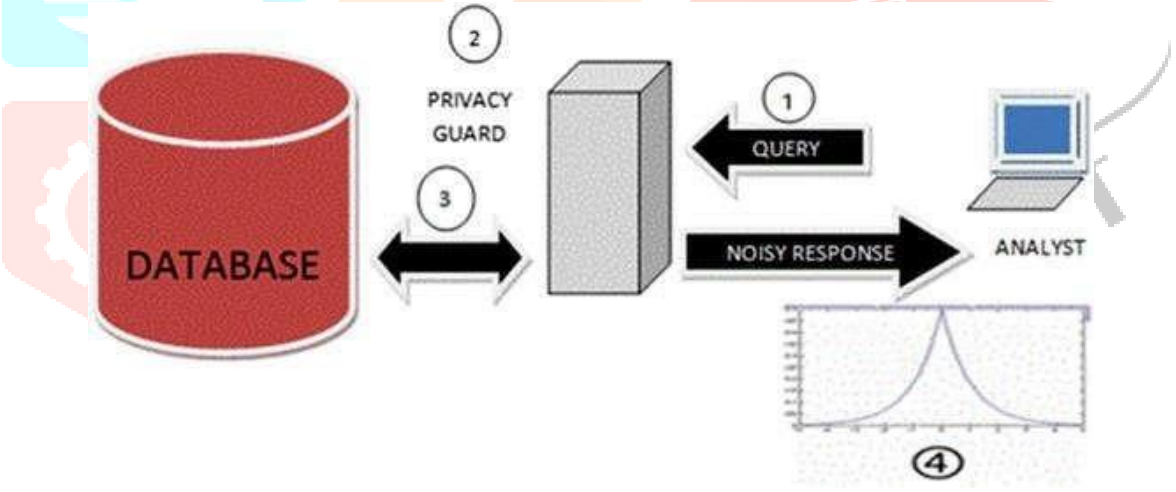


Fig 1: Differential privacy big data differential privacy (DP) as a solution to privacy- preserving in big data

Differential privacy big data differential privacy (DP) as a solution to privacy- preserving in big data is shown

Step 1 The analyst can make a query to the database through this intermediary privacy guard.

Step 2 The privacy guard takes the query from the analyst and evaluates this query and other earlier queries for the privacy risk. After evaluation of privacy risk.

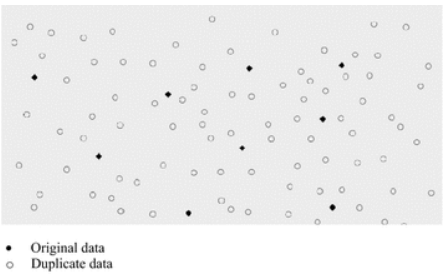
Step 3 The privacy guard then gets the answer from the database.

Step 4 Add some distortion to it according to the evaluated privacy risk and finally provide it to the analyst. The amount of distortion added to the pure data is proportional to the evaluated privacy risk. If the privacy risk is low, distortion added is small enough so that it do not affect the quality of answer, but large enough that they protect the individual privacy of database. But if the privacy risk is high then more distortion is added.

Identity based anonymization These techniques encountered issues when successfully combined anonymization, privacy protection, and big data techniques to analyse usage data while protecting

the identities of users. Intel Human Factors Engineering team wanted to use web page access logs and big data tools to enhance convenience of Intel's heavily used internal web portal. To protect Intel employees' privacy, they were required to remove personally identifying information (PII) from the portal's usage log repository but in a way that did not influence the utilization of big data tools to do analysis or the ability to re-identify a log entry in order to investigate unusual behaviour. Cloud computing is a type of large-scale distributed computing paradigms which has become a driving force for Information and Communications Technology over the past several years, due to its innovative and promising vision. It provides the possibility of improving IT systems management and is changing the way in which hardware and software are designed, purchased, and utilized. Cloud storage service brings significant benefits to data owners, say, (1) reducing cloud users' burden of storage management and equipment maintenance, (2) avoiding investing a large amount of hardware and software, (3) enabling the data access independent of geographical position, (4) accessing data at any time and from anywhere .

To meet these objectives, Intel created an open architecture for anonymization that allowed a variety of tools to be utilized for both de-identifying and re-identifying web log records. In the process of implementing architecture, found that enterprise data has properties different from the standard examples in anonymization literature . This concept showed that big data techniques could yield benefits in the enterprise environment even when working on anonymized data. Intel also found that despite masking obvious Personal Identification Information like usernames and IP addresses, the anonymized data was defenceless against correlation attacks. They explored the trade-offs of correcting these vulnerabilities and found that User Agent (Browser/OS) information strongly correlates to individual users. This is a case study of anonymization implementation in an enterprise, describing requirements, implementation, and experiences encountered when utilizing anonymization to protect privacy in enterprise data analysed using big data techniques. This investigation of the quality of anonymization used k- anonymity based metrics. Intel used Hadoop to analyse the anonymized data and acquire valuable results for the Human Factors analysts . At the same time, learned that anonymization needs to be more than simply masking or generalizing certain fields—anonymized datasets need to be carefully analysed to determine whether they are vulnerable to attack. Privacy preserving Apriori algorithm in MapReduce framework Hiding a needle in a haystack Existing privacy-preserving association rule algorithms modify original transaction data through the noise addition. However, this work maintained the original transaction in the noised transaction in light of the fact that the goal is to prevent data utility deterioration while prevention the privacy violation. Therefore, the possibility that an untrusted cloud service provider infers the real frequent item set remains in the method . Despite the risk of association rule leakage, provide enough privacy protection because this privacy-preserving algorithm is based on "hiding a needle in a haystack" concept. This concept is based on the idea that detecting a rare class of data, such as the needles, is hard to find in a haystack, such as a large size of data, Therefore, ought to consider a trade-off between problems would be easier to resolve with the use of the Hadoop framework in a cloud environment. In the dark diamond dots are original association rule and the empty circles are noised association rule. Original rules are hard to be revealed because there are too many noised association rules



Hiding a needle in a haystack Mechanism of hiding a needle in a haystack is shown In Fig. 2, the service provider adds a dummy item as noise to the original transaction data collected by the data provider. Subsequently, a unique code is assigned to the dummy and the original items. The service provider maintains the code information to filter out the dummy item after the extraction of frequent item set by an external cloud platform. Apriori algorithm is performed by the external cloud platform using data which is sent by the service provider. The external cloud platform returns the frequent item

set and support value to the service provider. The service provider filters the frequent item set that is affected by the dummy item using a code to extract the correct association rule using frequent item set without the dummy item. The process of extraction association rule is not a burden to the service provider, considering that the amount of calculation required for extracting the association rule is not much.

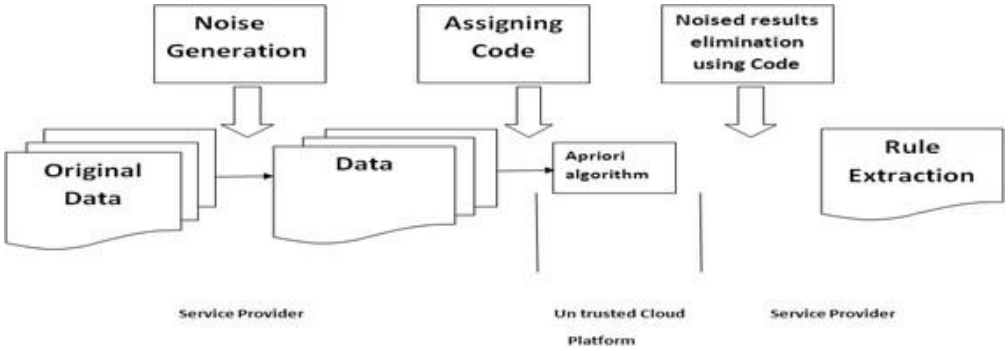


Fig 2 Hiding a needle in a haystack Mechanism of hiding a needle in a haystack

Overview of the process of association rule mining the service provider adds a dummy item as noise to the original transaction data collected by the data provider Privacy-preserving big data publishing The publication and dissemination of raw data are crucial components in commercial, academic, and medical applications with an increasing number of open platforms, such as social networks and mobile devices from which data might be gathered, the volume of such data has also increased over time . Privacy- preserving models broadly fall into two different settings, which are referred to as input and output privacy. In input privacy, the primary concern is publishing anonymized data with models such as k- anonymity and l-diversity. In output privacy, generally interest is in problems such as association rule hiding and query auditing where the output of different data mining algorithms is perturbed or audited in order to preserve privacy. Much of the work in privacy has been focused on the quality of privacy preservation (vulnerability quantification) and the utility of the published data. The solution is to just divide the data into smaller parts (fragments) and anonymize each part independently .

Despite the fact that k-anonymity can prevent identity attacks, it fails to protect from attribute disclosure attacks because of the lack of diversity in the sensitive attribute within the equivalence class. The l-diversity model mandates that each equivalence class must have at least l well- represented sensitive values. It is common for large data sets to be processed with distributed platforms such as the MapReduce framework in order to distribute a costly process among multiple nodes and accomplish considerable performance improvement. Therefore, in order to resolve the inefficiency, improvements of privacy models are introduced.

Trust evaluation plays an important role in trust management. It is a technical approach of representing trust for digital processing, in which the factors influencing trust are evaluated based on evidence data to get a continuous or discrete number, referred to as a trust value. It propose two schemes to preserve privacy in trust evaluation. To reduce the communication and computation costs, propose to introduce two servers to realize the privacy preservation and evaluation result sharing among various requestors. Consider a scenario with two independent service parties that do not collude with each other due to their business incentives. One is an authorized proxy (AP) that is responsible for access control and management of aggregated evidence to enhance the privacy of entities being evaluated. The other is an evaluation party (EP) (e.g., offered by a cloud service provider) that processes the data collected from a number of trust evidence providers. The EP processes the collected data in an encrypted form and produces an encrypted trust pre-evaluation result. When a user requests the pre-evaluation result from EP, the EP first checks the user's access eligibility with AP. If the check is positive, the AP re-encrypts the pre evaluation result that can be decrypted by the requester (Scheme 1) or there is an additional step involving the EP that prevents the AP from obtaining the plain pre-evaluation result while still allowing decryption of the pre-evaluation result by the requester (Scheme 2) .

Improvement of k-anonymity and l-diversity privacy model MapReduce-based anonymization

For efficient data processing MapReduce framework is proposed. Larger data sets are handled with large and distributed MapReduce like frameworks. The data is split into equal sized chunks which are then fed to separate mapper. The mappers process its chunks and provide pairs as outputs. The pairs

having the same key are transferred by the framework to one reducer. The reducer output sets are then used to produce the final result .

K-anonymity with MapReduce Since the data is automatically split by the MapReduce framework, the k- anonymization algorithm must be insensitive to data distribution across mappers. Our MapReduce based algorithm is reminiscent of the Mondrian algorithm. For better generality and more importantly, reducing the required iterations, each equivalence class is split into (at most) q equivalence classes in each iteration, rather than only two . MapReduce-based l-diversity The extension of the privacy model from k-anonymity to l-diversity requires the integration of sensitive values into either the output keys or values of the mapper. Thus, pairs which are generated by mappers and combiners need to be appropriately modified. Unlike the mapper in k-anonymity, the mapper in l-diversity, receives both quasi-identifiers and the sensitive attribute as input. Fast anonymization of big data streams Big data associated with time stamp is called big data stream. Sensor data, call centre records, click streams, and health- care data are examples of big data streams. Quality of service (QoS) parameters such as end-to-end delay, accuracy, and real-time processing are some constraints of big data stream processing. The most pre-requirement of big data stream mining in applications such as health-care is privacy preserving . One of the common approaches to anonymize static data is k-anonymity. This approach is not directly applicable for the big data streams anonymization. The reasons are as follows:

1. Unlike static data, data streams need real-time processing and the existing k- anonymity approaches are NP-hard, as proved.
2. For the existing static k-anonymization algorithms to reduce information loss, data must be repeatedly scanned during the anonymization procedure. The same process is impossible in data streams processing.
3. The scales of data streams that need to be anonymized in some applications are increasing tremendously.
4. Data streams have become so large that anonymizing them is becoming a challenge for existing anonymization algorithms.
5. To cope with the first and second aforementioned challenges, FADS algorithm was chosen. This algorithm is the best choice for data stream anonymization. But it has two main drawbacks:

1.The FADS algorithm handles tuples sequentially so is not suitable for big data stream.Some tuples may remain in the system for quite a while and are discharged when a specified threshold comes to an end. This work provided three contributions. First, utilizing parallelism to expand the effectiveness of FADS algorithm and make it applicable for big data stream anonymization. Second, proposal of a simple proactive heuristic estimated round-time to prevent publishing of a tuple after its expiration. Third, illustrating (through experimental results) that FAST is more efficient and effective over FADS and other existing algorithm while it noticeably diminishes the information loss and cost metric during anonymization process.

Proactive heuristic

In FADS, a new parameter is considered that represented the maximum delay that is tolerable for an application. This parameter is called expiration-time. To avert a tuple be published when its expiration-time passed, a simple heuristic estimated-round-time is defined. In FADS, there is no check for whether a tuple can remain more in the system or not. As a result, some tuples are published after expiration. This issue is violated the real time condition of a data stream application and also increase cost metric notably.

Privacy and security aspects healthcare in big data

The new wave of digitizing medical records has seen a paradigm shift in the healthcare industry. As a result, healthcare industry is witnessing an increase in sheer volume of data in terms of complexity, diversity and timeliness . The term “big data” refers to the agglomeration of large and complex data sets, which exceeds existing computational, storage and communication capabilities of conventional methods or systems. In healthcare, several factors provide the necessary impetus to harness the power of big data . The harnessing the power of big data analysis and genomic research with real-time access to patient records could allow doctors to make informed decisions on treatments . Big data will compel insurers to reassess their predictive models. The real-time remote monitoring of vital signs through embedded sensors (attached to patients) allows health care providers to be alerted

in case of an anomaly. Healthcare digitization with integrated analytics is one of the next big waves in healthcare Information Technology (IT) with Electronic Health Records (EHRs) being a crucial building block for this vision. With the introduction of HER incentive programs , healthcare organizations recognized EHR's value proposition to facilitate better access to complete, accurate and sharable healthcare data, that eventually lead to improved patient care. With the ever-changing risk environment and introduction of new emerging threats and vulnerabilities, security violations are expected to grow in the coming years.

Big data presented a comprehensive survey of different tools and techniques used in Pervasive healthcare in a disease- specific manner. It covered the major diseases and disorders that can be quickly detected and treated with the use of technology, such as fatal and non-fatal falls, Parkinson's disease, cardio-vascular disorders, stress, etc. We have discussed different pervasive healthcare techniques available to address those diseases and many other permanent handicaps, like blindness, motor disabilities, paralysis, etc. Moreover,a plethora of commercially available pervasive healthcare products. It provides understanding of the various aspects of pervasive healthcare with respect to different diseases.

Adoption of big data in healthcare significantly increases security and patient privacy concerns. At the outset, patient information is stored in data centres with varying levels of security. Traditional security solutions cannot be directly applied to large and inherently diverse data sets. With the increase in popularity of healthcare cloud solutions, complexity in securing massive distributed Software as a Service (SaaS) solutions increases with varying data sources and formats. Hence, big data governance is necessary prior to exposing data to analytics.

Data governance

1. As the healthcare industry moves towards a value-based business model leveraging healthcare analytics, data governance will be the first step in regulating and managing healthcare data.
   2.The goal is to have a common data representation that encompasses industry standards and local and regional standards.
2. Data generated by BSN is diverse in nature and would require normalization, standardization and governance prior to analysis.

Real-time security analytics

1. Analysing security risks and predicting threat sources in real-time is of utmost need in the burgeoning healthcare industry.
2. Healthcare industry is witnessing a deluge of sophisticated attacks ranging from Distributed Denial of Service (DDoS) to stealthy malware.
3. Healthcare industry leverages on emerging big data technologies to make better-informed decisions, security analytics will be at the core of any design for the cloud based SaaS solution hosting protected health information (PHI).

Privacy-preserving analytics

1. Invasion of patient privacy is a growing concern in the domain of big data analytics.
2. Privacy-preserving encryption schemes that allow running prediction algorithms on encrypted data while protecting the identity of a patient is essential for driving healthcare analytics .

Data quality

1. Health data is usually collected from different sources with totally different set- ups and database designs which makes the data complex, dirty, with a lot of missing data, and different coding standards for the same fields.
2. Problematic handwritings are no more applicable in EHR systems, the data collected via these systems are not mainly gathered for analytical purposes and contain many issues—missing data, incorrectness, miscoding—due to clinicians' workloads, not user friendly user interfaces, and no validity checks by humans .

Data sharing and privacy

1. The health data contains personal health information (PHI), there will be legal difficulties in accessing the data due to the risk of invading the privacy.
2. Health data can be anonymized using masking and de-identification techniques, and be disclosed to the researchers based on a legal data sharing agreement .
3. The data gets anonymized so much with the aim of protecting the privacy, on the other hand it will lose its quality and would not be useful for analysis anymore And coming up with a balance between the privacy-protection elements (anonymization, sharing agreement, and security controls) is essential to be able to access a data that is usable for analytics.

Relying on predictive models

1. It should not be unrealistic expectations from the constructed data mining models. Every model has an accuracy.
2. It is important to consider that it would be dangerous to only rely on the predictive models when making critical decisions that directly affects the patient's life, and this should not even be expected from the predictive model.

Variety of methods and complex math's

1. The underlying math of almost all data mining techniques is complex and not very easily understandable for non- technical fellows, thus, clinicians and epidemiologists have usually preferred to continue working with traditional statistics methods.
2. It is essential for the data analyst to be familiar with the different techniques, and also the different accuracy measurements to apply multiple techniques when analyzing a specific dataset.

CONCLUSION

This Paper reveals that safeguarding and preserving the privacy of big data is the top priority in order to protect data from harmful assaults and to ensure the safety of data to prevent it from getting into the wrong hands, regardless of how sophisticated the underlying big data technology may be. De-identification and encryption are only two examples of the various methods available for shielding big data from potential threats. Data Cryptography, End Point Filtration, etc. are various technologies used to address data security and privacy, each using a unique set of methods and algorithms. However, despite the abundance of security data, protecting information from assaults remains challenging due to data's amount, diversity, velocity, and accessibility. This highlights the necessity for cutting-edge methods of addressing Big Data Privacy and Security to finally put a stop to such scams.

REFERENCES

1. Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. Journal of Big Data, 3(1), 1-25.
2. Pramanik, M. I., Lau, R.Y., Hossain, M. S., Rahoman, M. M., Debnath, S. K., Rashed, M. G., & Uddin, M.Z. (2021). Privacy preserving big data analytics: A critical analysis of state-of-the-art. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(1), e1387.
3. Gao, Weichao & Yu, Wei & Liang, Fan & Hatcher, William & Lu, Chao. (2018). Privacy-Preserving Auction for Big Data Trading Using Homomorphic Encryption. IEEE Transactions on Network Science and Engineering. PP. 1-1. 10.1109/TNSE.2018.2846736.
4. Iezzi, M. (2020, December). Practical privacy-preserving data science with homomorphic encryption: an overview. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 3979- 3988). IEEE.