Secure EMR Classification and Deduplication using MapReduce

USHARANI A V, GIRIJA ATTIGERI
Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India 576104
Corresponding author: Girija Attigeri (e-mail: ga.research10@gmail.com).

ABSTRACT Healthcare providers generate huge amount of data every day through registration, lab results, prescriptions, and others. This is stored in the form of Electronic Medical Records (EMR) in a central repository. A medical record data is very huge, difficult to read and understand. To give an insight to the professionals in analysing the different domains a patient belongs to, it is necessary to get pointers to a file before classifying it to a particular department for further analysis. This study provides an EMR processing system to automatically classify EMRs based on the important medical terms using TF-IDF and topic modelling. Automatic Classification of EMRs help the healthcare professionals in taking accurate decisions, providing efficient service, and improves the time taken for processing huge amount of data and in better organizing of patient. The data stored on the cloud may contain duplicate copies of EMR on several storage systems at file level thus increasing the network bandwidth, cost, and consuming storage space. Hence, a deduplication mechanism is required to avoid or reduce the data redundancy. Adapting cloud computing for healthcare systems necessitates sharing patient data with cloud service providers, which creates security concerns as the data may contain diagnosis, medication, laboratory results and medical claims. The main aim of this work is to classify the EMRs as per the specialization using KNN algorithm, optimize storage using deduplication and protect the data using DNA encryption algorithm before uploading to Hadoop. Data redundancy is taken care by implementing deduplication techniques using MD5 hashing.

INDEX TERMS Classification, Clustering, Deduplication, DNA Encryption, Electronic medical records, Hadoop, Map Reduce

I. INTRODUCTION

URING the most recent decade, there has been a huge appropriation of cloud-based Electronic Health Records (EHRs) and Electronic Medical Records (EMRs) by various clinical bodies. As indicated by National Health Record Survey 2017, the level of office-based doctors using any EHR or EMR framework is 85.9%. There has been a consistent development in the reception of public electronic well being record frameworks in recent years and 46% worldwide expansion in the previous five years as expressed by WHO.

At first, EHRs were intended for filing and getting sorted out patients’ records. With time, EHR driven information bases turned out to be more extensive, dynamic, and interconnected. Health experts (physicians or nurses), health facilities (clinics, hospitals for providing medications and other diagnosis or treatment technology), and a funding institution supporting the first two are the key components of a healthcare system. Various levels of healthcare are required depending on the severity of the condition. Professionals use it for primary care, acute care that requires experienced professionals (secondary care), advanced medical research and treatment (tertiary care), and extremely rare diagnostic or surgical operations (quaternary care). Health professionals oversee many types of information at all these levels, including the patient’s medical history (diagnosis and prescription data), medical and clinical data (such as data from imaging and laboratory procedures), and other private or personal medical data. A large volume of patient data has been gathered and is now accessible online, including lab tests, drugs, disease history, and treatment outcomes. Advanced clinical data analysis and data management have been developed to incorporate patient data into a database of rich longitudinal patient profiles. EMR/EHRs open previously unimagined possibilities for cohort-wide research and knowledge discovery. They are vital data sources for developing predictive models for disease diagnosis and prognosis, allowing personalised medicine to
flourish. Hence, this work uses Latent Dirichlet Allocation (LDA) topic modelling to find relevant medical topics in an EMR. LDA is used in various domains of natural processing and data mining. Each document is made up of different terms, and each topic has its own set of words. LDA's task is to identify which topics a document belongs to based on the set of words it contains.

The computing ability needed to process massive number of varieties of healthcare data rapidly can burden a system or network of servers. Thus, organizations need to adopt cloud computing to achieve sufficient processing power and speed needed to handle big data tasks. Cloud computing offers a new model for enhancing healthcare delivery and expanding medical organisations’ business flexibility, allowing them to operate with greater efficiency, cost-effectiveness, and flexibility. Hence, numerous servers may often focus on Hadoop and Apache Spark’s technologies, where the work is distributed among several nodes in a clustered way. To achieve such pace cost-effectively is also a challenge. Cloud computing offers users endless storage space, processing power, apps, servers, networking, and ease of access to data always. Cloud computing’s primary aim is to provide services focused on virtual and decentralized computer technology.

As the use of digital medical records increases, the amount of duplicate data also increases leading to data redundancy. Redundant data need to be reduced by removing the duplicate copies of data or by avoiding the upload of redundant data. Replicate data has been identified to the point that 50% of the cloud space is filled. The expense of managing similar data sets is excessive, almost eight times higher than processing the native cloud data. Reducing management spending is very important with the rise in data submitted to the cloud and improving the use of storage. Enhancing storage performance and management expense optimization is, therefore, an urgent issue to be addressed. The cloud service providers should implement data deduplication techniques to deal with the above issue, remove duplicate data copies, optimise the use of storage, and reduce processing expenses. This can be done using various deduplication techniques. Deduplication is the method of removing duplicate data copies by means of a deduplication scanning procedure to store unique copies and then support all approved users. Deduplication is a critical component of any storage system.

Data stored in the cloud may contain personal, private, or confidential information, such as medical records, that must be protected against disclosure, breach, or misuse. Concerns about data jurisdiction, security, privacy, and compliance are impeding adoption by healthcare companies around the world. EMRs hold personal information related to patient’s medical history, treatment, and other details such as medical claims, etc. Medical organizations share and manage this information for various purposes like billing, research, and storage. Hence, security is very critical in health systems to provide protection to patient data.

EMR is a digital record of patient information such as vital signs, medical history, treatment, diagnosis, scanning, radiology and other laboratory results, billing and medical claims maintained by healthcare organisations for improving the operations and coordination for efficient patient care. EMRs enable the doctors and physicians to extract useful information regarding medications and treatments, reduce errors and to improve health care. Follow-up instructions, scheduling appointments, access to lab results, and reminders for self-care can be provided to patients and thus helps in lifestyle changes to improve overall health. It also improves the quality of operation in hospitals, improves the documentation, enables the tracking and searching of patient related information in an organized and efficient manner.

Research work has following contributions:
1) Finding relevant medical topics for classification from EMR.
2) To classify the EMR’s using KNN MapReduce.
3) Optimize data storage on Hadoop using deduplication techniques.
4) To provide data protection using DNA encryption.

The rest of the paper is organized as follows. The first section gives a literature review on topics like Topic modelling, Hadoop, map reduce, different ways of extracting information from EMR, classification and clustering algorithms. The next section gives a detailed explanation about the proposed system. It explains the architecture of the model, KNN map reduce, DNA encryption and decryption algorithms and deduplication technique. The fourth section discusses about the results and its importance. The last section deals with the conclusion of the work and future scope.

II. RELATED WORK

Topic discovery or modelling plays a very important role in feature extraction for different domains. Chen Bingyang et al [11], studied the emotions of tourist visiting a scenic spot using LDA topic modelling to cluster the tourism dataset. An emotional dictionary is constructed separately for online reviews and microblogging texts. Initially Naives Bayes algorithm is used to for constructing a clear emotional parity and the LDA theme model is added to refine the errors during reclassification of another part. A Tourist corpus was selected from microblog and travel website. At last, the classification result of the two classifiers is combined to get the emotions of tourists.

A novel domain-oriented topic discovery was done by Xiaofeng lu et al [2], to identify emerging cyber threat topics in the domain in real life to analyse open-source platforms and blogs. This Feature Extraction and Topic clustering (FETC) applied to cyber security data threat considers location of word and parts of speech. They first used web crawlers to extract data from specific labels such as vulnerability and malware. Then, the article title, body text and time are extracted and stored in a database. The authors then offer three improved feature extraction methods: a better keyword feature extraction approach, a subject word feature extraction method, and an entity feature extraction method.
The obtained features are combined using feature fusion technology, and the feature vector of the article is created for use as the topic clustering module’s input. Finally, an improved hierarchical clustering algorithm is implemented to cluster the feature vectors of articles in each period to identify emerging or historical topics in real time.

A knowledge base was created by Girija Attigeri et al [3], for fraud detection in news related articles. Articles from news websites were collected and pre-processed by removing stop words and punctuations. The words with higher TF-IDF scores provide important insight about the fraud. The outcome of LDA indicates any suspicious activity that may lead fraudulent incidents. The topics derived from LDA are used to create a knowledge base ontology using which may be used for the classifying the transactions as legal or fraud.

To ease the problem of sparsity in short texts, a sentence based LDA was set up by Fan Zhang et al [4]. A single topic is used for generating the short text words. To do this, a word embedding model with replacement is used to recognise the relationship between the words. This relationship plays an important role in determining the relevance or importance of a word. In practice, a health-care decision-making system assesses whether a patient has an illness based on a set of symptom indexes and physical characteristics; disease diagnosis is, in essence, a classification problem. Data mining technology can assist healthcare by grouping patients with similar ailments so that healthcare providers can give them with better treatment [5].

A. EMR DATA EXTRACTION AND DEDUPLICATION

Due to the expanding number of health records and complexities associated with the health business, several challenges will be met by the healthcare data managers. The burden on e-health data managers can be reduced by automating their work. Heath [6], described the distribution of scanned documents at their institution. The author developed a system that categorizes the EMRs based on the design. The author evaluated the text classification models that were trained on documents through optical character recognition. The author made use of previously manually classified documents to determine the accuracy of their system. Thus, their system could classify the documents based on the clinical relevancy. Thus, using machine learning and OCR approach they were able to classify the documents into clinically relevant documents.

Policy-makers and other key stakeholders will use the key organizational, human, and technical factors described in this analysis to make evidence-based decisions during the implementation of a completely inter-operable EHR across primary, secondary, and long-term care. Orna Fennelly et al [7], aimed at identifying the key factors that impact on successful implementation of EHR at different levels of e-healthcare. They concluded that the efficient implementation of these variables, however, requires careful consideration of the contextual influences. End-users, current technical requirements, and policies, as well as developments in technology and research in the field, will all influence how these factors communicate dynamically during EHR implementation and affect performance. Isabel Segura-Bedmar et al [8], explored different techniques that automatically classify EMRs. They also compared various ways of representing EMR and applied some efficient text classifiers to identify clinical records describing anaphylaxis (an allergic reaction) cases. Their experiments revealed that the prediction of anaphylaxis cases is a linear problem that thus they can be efficiently solved by using linear classifiers such as Linear SVM or Logistic Regression and the BoW (Bag of Words) representation. A simple CNN architecture achieves the top F1, but it takes about 50 minutes to train the network. High F1 (91–94%) and low training times are given by k-NN, Linear SVM, and Logistic Regression (less than 20 s). The fastest algorithm is k-NN, which takes less than a second to learn.

Different analytic tools can be adapted in various stages of implementing EHR. Fadoua Khennou et al [9] studied the challenges in implementing the EHRs and their association with different health systems. They could successfully implement an OpenEHR model to study the intermediate steps involved in it. A processing technique that involved in depth study of different steps of processing EMRs like data cleaning, integrating, transforming and privacy protection was proposed by Wencheng Sun et al [10], reviewed different approaches of extracting information from EMRs that involved structured and unstructured data. Unstructured data in clinical notes can be processed to help doctors arrive at certain important decisions. Wencheng Sun et al [11], have discussed the current state of data mining technology research in EMR and the obstacles that EMR research is currently facing. The Electronic Medical Record (EMR) is used by medical institutions to keep track of a range of medical operations, including diagnostic details (diagnosis codes), procedures performed (procedure codes), and admission records. Due to the growing social value, it creates, it has become a hotspot for experts and academics.

The inherent characteristics of electronic medical records were examined by Yuan Zhang [12] from real-world electronic health (eHealth) systems, finding that, first, multiple patients produce many duplicate EMRs, and second, cross-patient duplicate EMRs are produced in large numbers only when patients visit doctors in the same department. They proposed effective and stable encrypted EMRs deduplication scheme for cloud-assisted e-Health systems (HealthDep). HealthDep enables the cloud server to effectively conduct EMR deduplication, resulting in a 65 percent reduction in storage costs while maintaining EMR confidentiality. Parallelism was implemented by Dongzhan Zhang et al [13], by using MapReduce technique and HBase to accelerate the deduplication process. However, one among the critical challenges faced by file service providers is that they often need to contain duplicate copies of file contents as data are being accumulated. A new file aggregation scheme supported MapReduce is proposed and therefore the recent SHA-3 standard Keccak is employed for hash computation. Shahid Mu-
nir Shah et al [14], have analysed various ways of mining data from EHRs. Healthcare providers produce massive quantities of clinical data on a regular basis in today’s technological age. As a central data repository for hospitals, generated clinical data is stored digitally in the form of an Electronic Health Record (EHR). The information stored in EHR is used for a variety of secondary purposes, including clinical testing, automated disease monitoring, and clinical audits for quality improvement. Individuals’ privacy may be jeopardised if confidential personal information stored in EHR is leaked or revealed to the public. Data breaches may result in financial damages, and if a person’s medical condition is revealed in public, he or she can face social boycott. Different privacy laws exist to safeguard patients’ personal data from certain risks, such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and My Health Record (MHR). However, as state-of-the-art techniques in Machine Learning (ML), Data Analytics (DA), and hacking continue to evolve, it is becoming very difficult to fully protect a patient’s privacy. They have thoroughly analysed various uses of EHR in the article with the aim of highlighting how the secondary analysis impact patients’ privacy. A list of the benefits, implementations, research methods, and challenges of Big Data for healthcare were reviewed by Liang Hong et al [15]. Heterogeneity, incompleteness, timeliness and persistence, anonymity, and ownership are all characteristics of Big Data in health care. These characteristics present several difficulties in terms of data storage, mining, and sharing to advance health-related science. To address these issues, research methods based on Big Data in health care must be created, as well as laws and regulations governing the use of Big Data in health care. From the standpoint of the patient, Big Data analysis may result in better care and lower costs. The government, hospitals, and research institutions, in addition to patients, could benefit from bigdata in healthcare.

Caifeng Zhang et al [16], have emphasized on efficient adoption of EHRs in medical field. The China government issued the Electronic Health Record Architecture and Data Standard in 2009 as a guide for hospitals with the aim of enhancing treatment quality through the meaningful use of electronic health records (EHRs). The Chinese EHR Standard defines EHRs as “A full set of digital clinical information recording the clinical treatment given to a person”. EHRs have been recommended for over two decades to improve the quality and efficacy of healthcare services, but these do not guarantee that implementing the EHRs scheme would result in such benefits. To know the payback, healthcare professionals must make the EHR a part of their everyday work routine. As a result, the Health Information Technology for Economic and Clinical Health (HITECH) Act establishes “meaningful use” of electronic health records as a target for adoption. Act’s key goal is to develop meaningful and usable digital medical records, including electronic health records (EHRs) entry and storage, as well as to maximise the use of EHRs.

Using deterministic record linkage algorithms, a stable protocol for deduplication of horizontally partitioned data sets was developed by Kassaye Yitbarek Yigzaw et al [17]. They have found various techniques for computing statistics on distributed data sets without disclosing any personal data other than the statistical results. In a distributed data set, however, duplicate records can lead to incorrect statistical results. As a result, safe deduplication is an essential pre-processing step for improving the accuracy of statistical analysis of a distributed data set. They targeted on the reuse of health records horizontally partitioned among records custodians, such that every records custodian presents the equal attributes for a fixed of patients. Reusing records from more than one records custodian presents sufficient patients who satisfy the inclusion standards of a selected study. The variety of patients at a single records custodian can also additionally offer inadequate statistical power, especially for research on uncommon exposures or outcomes. When records are acquired through more than one source, the records of a heterogeneous blend of sufferers may be reused.

B. HADOOP MAPREDUCE

Due to the rising number of health records, there will be many problems to overcome. And the problems that come with working in the health-care sector as a result, it is critical to convert the data size and quality into a crucial nominal value using all available solutions. In terms of safety and efficiency problems, detailed research need to be made. An attempt to meet the requirements of privacy, reliability, and accuracy separately was made by Jiawei Yuan et al [18]. They first proposed a privacy-preserving algorithm based on clustering using K-means for handling extensive dataset on public cloud domain. Centered on the hard problem of Learn with Error (LWE), a new encryption method was discovered, which achieved privacy-preserving calculation of data object similarity upon cipher-texts. With the help of their encryption program, they built the entire K-means clustering scheme in a privacy-safeguarding mode, wherein cloud servers having access to encrypted data performed all operations without any decryption. In this way, along with creating a new method in cloud computing, they also achieved better performance compared to the K-means clustering without security. Considering the support of large-scale datasets, they integrated the MapReduce system into the new architecture, and it was highly appropriate for parallelized environmental analysis for cloud computing.

An agglomerative fuzzy algorithm based on K-means with the MapReduce technique was implemented by Ruixin Zhang et al [19]. In this algorithm, an initial centre is selected to improve the convergence speed of the k-means algorithm. Then, to improve scalability for large datasets, a MapReduce implementation based on Apache Hadoop is presented. Though K-means is a simple algorithm, determining the initial centroid is very difficult and if the data set is very large all the variables have the same effect and it may lead to dimension trap. To solve this dichotomous method
is used in which two objects which are far are selected as initial centroids as explained by Qingyuan et al. [20]. Yurong Zhong et al. [21], focused on integrating both property and location of K-means clustering on Hadoop platform. The new cluster will be calculated based on Map function. They also analyzed the time complexity of the algorithm. A distributed environment and parallel processing of data was proposed by Wooyeol Kim et al. [22] to process the KNN joins using MapReduce for Bigdata. This KNN-MR algorithm identifies the data that need to be eliminated by using non-KNN points that utilizes vector projection pruning. This approach reduces the computational and network costs across the machines.

The problem of discovering knowledge and making decisions from rapidly rising voluminous big data is a difficult one, K. Sharmila et al. [23]. In developing countries like India, diabetes mellitus (DM) is a major health concern. The acute essence of diabetes mellitus is linked to long-term complications and various other ailments. Based on their findings, they propose an effective predictive approach, the MRK-SVM hybrid algorithm, for identifying diabetes types and predicting diabetes complications at an early stage. For big data analysis, a real-time dataset was generated by replicating records from five districts in Tamilnadu, and big data analysis was performed using Hadoop MapReduce, Spark, and in the cloud.

Srikanth Bethu et al. [24] have inferred that the pursuit of knowledge and its application in computationally intensive activities with a wide variety of applications, data mining plays a critical role. A theoretical and experimental comparison of the approaches for computing KNN on MapReduce is done. Each stage of data pre-processing, data partitioning, and computation is examined for load balancing, accuracy, and complexity. The experiment results are produced from a variety of datasets. Each dataset’s time and space complexity are examined on a regular basis, yielding new benefits and drawbacks that are addressed for each algorithm. Finally, they have discussed various ways of dealing with KNN-based problems in the context of Big Data MapReduce.

Shaobo DuJing li et al. [cite23] have proposed Hadoop-based parallel processing of an improved KNN text classification algorithm. They state that the network has become a significant platform for people to share information due to the rapid growth of mobile Internet. The study of text classification is useful in the real world. The Hadoop platform can be used to parallelize the KNN classification algorithm, which can easily and accurately classify text. However, when measuring the similarity or distance of samples, the KNN algorithm will increase as the sample data grows, increasing the algorithm time. The CLARA (Clustering Large Applications) clustering algorithm is used to take out samples in the dataset that have low similarity, and the sample distance measurement in the dataset is reduced. Subsequently, to identify the network public opinion data, parallel KNN MapReduce program is designed. The results of the experiments show that the improved parallel KNN algorithm improves text classification accuracy and speed.

Doreswamy et al. [25] have suggested a k-means clustering algorithm that scales up to a large dataset of about 8 million items. Each object is a five-attribute vector. To find the maximum value of inter-cluster density and the minimum value of intra-cluster measurements, inter and intra cluster measurements were computed. Dillon Chrimas et al. [26] developed a HBDA (Healthcare Bigdata) platform for health applications was built using a Hadoop/MapReduce framework. Only a few works have used big data tools for analyzing hospital patient data for healthcare applications. This solution removes the need to move data in and out of the storage system while parallelizing the computation. This problem is very relevant in healthcare due to the growing number of sensors and data produced. Furthermore, making a stand on using the Apache software needs usability goals-based end user computing of Big Data technology and leveraging existing tools from a data warehouse at a health authority. A stable medical big data ecosystem on top of the Hadoop big data platform to increase the intelligence of the medical system was implemented by Xiangfeng Zhang et al. [27]. It was created in response to the current security medical big data ecosystem’s increasingly serious trend. Authors emphasized utilizing block-chain as it has been a successful breakthrough point for innovation in medical data interchange since it is a distributed accounting system for multi-party maintenance and backup information security. The customized health information system for stroke has been built to deliver personalized care for patients and to promote the management of patients by medical professionals by using the advantages of the Hadoop big data platform.

C. DATA SECURITY

Prasanna Balaji Narasingapuram et al. [28], suggested a new form of cryptography to detect and remove malicious users to improve security at the user level to prevent the entry of malicious users into cloud applications. Various cryptographic approaches, algorithms, and strategies for user verification for data access have been proposed in their research. But malicious user activities are still growing day by day in the cloud. This approach focuses on eradicating malicious cloud user behavior. In-order to do this, the device uses the method of DNA cryptography to produce a strong user key and the decryption process for data encryption. The findings are analyzed with the method’s current outcomes and have shown that the proposed DNA outperforms other cryptographic techniques. In-order to reduce the burden on network bandwidth and improve storage R. Shiny Sharon et al. [29], proposed a deduplication technique. In terms of encryption and decryption, File Hierarchy-ABE is used, which has a low storage cost and computational complexity. Privacy preserving technique was also implemented for securing the PHI (Personal Health Information) such as the Layered Model of Access Structure, which addresses the issue of sharing multiple hierarchical files. In-order to improve the efficiency on the cloud, Huqi Zhao et al. [30] proposed that convergent encryption be applied to the medical cloud, with the
convergent key’s hash value being used as a repeat detection label and Bloom filter search to improve the medical cloud’s efficiency, and the practicality of medical data be improved by introducing fuzzy keywords to address the problem of resource waste caused by the repeated storage of medical data. At the same time, the identity token authorization access methodology was to be used to achieve multi-function access to the medical cloud based on the real application of diverse medical data.

III. METHODOLOGY

First the EMR is processed by removing the stop words, lemmatized by changing verbs in past and future tense to present tenses. Words are then stemmed by using corresponding root words. A dictionary containing the most frequent words with their number of appearances is created using gensim package. LDA model is then trained with the key phrases and their weights to obtain the topics related to medical terms along with their TF-IDF scores. TF-IDF scores indicate the significance of a term in a particular EMR. Thus, the knowledge obtained from this algorithm is used for classifying the EMRs to the most significant department. For the Map Reduce technique, the proposed work utilizes optimized KNN method on large Dataset on Hadoop. A distributed file system has the benefit of Hadoop and provides fast file access. The architecture diagram shown in Figure 1 depicts the different components of the proposed system. Database is used to store the hash codes of the blocks generated from MD5 and to store the Logical Block Addressing of respective files. Hadoop HDFS is used to store the blocks related to the user files. The file will be classified based on the resemblance to the trained dataset class labels.

Dataset : The input data will be the patient EMR (text document). It consists of patient’s history, detailed explanation consisting of the vital signs of patient along with the treatments suggested by the doctor. The vital signs or the medical terms related to a particular department play a major role in classifying the patient EMR. Before testing the system, it is trained with the data related to each department. The following departments related to medical field are considered: Cardiology, Physiology, Pathology, Paediatrics, Obstetrics, Histology, Gynaecology, Oncology, Neurology, Forensic, Dental, Digestive, Ophthalmology, Orthology, Respiratory, Surgery, Micrology, Otorhinolaryngology, Nephrology.

The system is first trained with datasets related to different medical departments. The file selected by the user must be saved in that category. The selected files will be split into blocks and hash code is generated using MD5 algorithm. If the hash tag matches, we will not upload the block to Hadoop, we will increase the number of instances of that block in the database table. Subsequently hash blocks compared with the current hash tag from the database. If the hash tag is not matched, we will add the information of the block hash to the database and upload the block to Hadoop. The technique of Logical Block Addressing (LBA) is used to define which blocks in a file are present. The process flow is shown in the Figure 1.

First, for each distinct class label (department) that is linked to medical knowledge, we upload a qualified data collection. Each keyword is associated with a weight based on the number of occurrences in the data set. The words in the user file are compared with the trained data set. Each class has a center data point, and the keyword is compared with the data point in the trained data. If it matches, the weight-age of the word is calculated using Euclidean distance. The keyword and its weight-age are stored as key value pair and used by the MapReduce algorithm to recalculate the weight-age of words.

Based on the combined weight of the words, and its distance with the user file, the file is given a class label. The class with maximum of matches will be chosen using the single nearest neighbour KNN classification. After that, the
file is stored in the respective class directory in the eclipse.

The blocks that are already stored in the database are not uploaded to the HDFS. After that, DNA encryption is applied to the blocks of data created to make it secure before storing it on the Hadoop HDFS. To write data to HDFS, the client must first contact with the name node to ensure that the file can be written, and the data node that will receive the file block. The client then sends the file in order, block by block, to the associated data node, and is responsible for ensuring that the data node receives each block. The file upload algorithm steps are explained in the algorithm 1.

Algorithm 1 File Upload Algorithm
1) Read File F from the perspective.
2) Based on the packet size file chunks will be formed
3) for I= 0 to N do Generate the hash code for each chunk using MD5 algorithm
4) Compare the hash code with the existing hash code in the database.
5) if exists then
   • get the block id of the identical hash code from the database.
   • increase the Instance by Map Reduce Technique (Mapping to the existing Block)
6) else
   • Insert the new hash code to the database and get the id of the inserted hash code for the LBA process.
   • Upload the block to the HDFS storage end if

When the file needs to be downloaded, Logical Block Address of the file is obtained from the file id as explained in the algorithm 2. The blocks are downloaded from the HDFS, and the encrypted blocks will get decrypt by using DNA algorithm then merge the blocks and give it to the user. Because of its reliability and simplicity, the k-Nearest Neighbour classifier is one of the most well-known data mining approaches. This model enables us to categorise many previously unseen cases (test instances) against a huge (training) dataset at the same time. To do so, the map phase will find the k-nearest neighbours in various data partitions. Following that, the reduce stage will compute the definite neighbors from the map phase’s list. After that, it will use the k-NN model’s typical majority voting technique to predict the resulting class. The numeric vector represents the input document data-set after pre-processing the data-set and the vectors are stored in the register. The number of the vector values is taken at random from the data-set. The below algorithm provides the implementation of the KNN Map-Reduce:

Algorithm 2 KNN Algorithm
1) Extract the keywords with their weightage and store it in an array.
2) Let N be the number of classifications and M the number of extracted words.
3) Initialize an array – weight[N]
4) For I=1 to M, J=1 to N Check the presence of K word in classification keywords,
5) if it is present, weight[J]=weight[J]+ K word Next J, Next I.
6) Fetch the next highest weight value and index, add the index in classification array.
7) W = W + Fetched class weight. If W >=Threshold, goto step 7
8) Print all the categories in the classification array.

The DNA carries the genetic instructions used in the growth, development, functioning and reproduction. Nucleotides are composed of DNA and RNA, which in turn are composed of four nucleotide groups: cytosine(C), guanine(G), adenine(A) or thymine(T), deoxyribose and phosphate. The hydrogen bonds and nucleobases form double-stranded DNA to store biological information according to the base pair law. As a key for the encryption algorithm, this knowledge is used. 1021 bases, equal to 108 terabytes, are contained in a gram of DNA. One gram of DNA contains 1021 bases of DNA, corresponding to 108 TB of data. Hence, in a few milligrams, all the data in the world can be stored. Cryptography based on DNA is the technique of using biological structure to hide data and details. The DNA algorithm works as explained in the algorithm 4. The decryption algorithm is reverse of encryption.
Algorithm 4 DNA Algorithm

1) Get the block to be encrypted.
2) Change it to ASCII value
3) Convert it to binary (0’s and 1’s)
4) Change it to DNA code.
5) A random key is generated between 1-256
6) 256 index values are created using permutation of 4 letters A,T,C,G (stored in a table )
7) Change it to DNA code
8) convert to binary and to corresponding numbers
9) Numbers - output

IV. RESULTS

The topics obtained from the LDA algorithm along with their TF-IDF scores is shown in Figure 2. This score indicates the importance of a word in the EMR document. The first topic 0 indicates the words related to cardiac. The next topic 1 recognizes the words related to operation, topics 2 and 3 dental and respiratory topics.

The performance of the system is evaluated by calculating the accuracy, F1 score, Kappa, Recall and Precision values depicted in Figure 3. The classification algorithm was evaluated for 5 classes and the true labels vs predicted labels are shown in Figure 3. Accuracy is the number of EMRs correctly classified to a particular department to the total number of EMRs. F1 score represents the harmonic mean of precision and recall. The accuracy is above 85% for all the classes and the F1 score is above 69%. The graph shows a good recall and precision values. The Kappa value generated is more than 0.60 indicating moderate inter reliability. The Figure 4 shows calculations related to kappa statistics.

The Figure 5 and explains the sensitivity and specificity ratios of the model. The positive and negative RCE indicate how likely an EMR belongs to a particular department. Positivity tells us how much to increase the probability of belonging to a particular department and negativity explains how to much probability to decrease. The values of positive and negative likelihood ratios are shown in the Figure 6.

A. DE Duplication TESTCASE

In order to test deduplication, two EMRs named Patient1 and Patient2 with same contents were uploaded. The first file Patient1 was uploaded successfully. But when the second file with the same content was given for upload, it was not allowed to upload on HDFS, and the logical addressing of the blocks was stored in the database as shown in the table 1. Hence, by not allowing the upload of similar files,
of terabytes of data, Hadoop technology is successful in tackling the above issues faced by the healthcare business. In this work, a Map Reduce based KNN classification is proposed to classify large amounts of patient EMR into different departments of medical using the vital signs of patient as data points. To do so, the map phase will find the k-nearest neighbors in various data partitions. Following that, the reduce stage will compute the definite neighbors from the map phase’s list. The suggested approach enables the k-Nearest neighbor classifier to classify the patient EMR to different departments of medical effectively. The DNA algorithm provides the necessary security to the data and deduplication reduces the storage cost and the number of transactions to the cloud is reduced by uploading the blocks that are unique.

V. CONCLUSION

In today’s healthcare systems, digitization of all clinical evaluations and medical information has become a regular and widely accepted procedure. The huge amount of data that gets generated through various devices, need to be analysed to provide efficient service by reducing the errors. Because the MapReduce engine and HDFS can process thousands

of terabytes of data, Hadoop technology is successful in tackling the above issues faced by the healthcare business. In this work, a Map Reduce based KNN classification is proposed to classify large amounts of patient EMR into different departments of medical using the vital signs of patient as data points. To do so, the map phase will find the k-nearest neighbors in various data partitions. Following that, the reduce stage will compute the definite neighbors from the map phase’s list. The suggested approach enables the k-Nearest neighbor classifier to classify the patient EMR to different departments of medical effectively. The DNA algorithm provides the necessary security to the data and deduplication reduces the storage cost and the number of transactions to the cloud is reduced by uploading the blocks that are unique.

V. CONCLUSION

In today’s healthcare systems, digitization of all clinical evaluations and medical information has become a regular and widely accepted procedure. The huge amount of data that gets generated through various devices, need to be analysed to provide efficient service by reducing the errors. Because the MapReduce engine and HDFS can process thousands
[18] J. Yuan and Y. Tian, “Practical privacy-preserving mapreduce based k-means clustering over large-scale dataset,” IEEE transactions on cloud computing, pp. 568–579, 2017.

[19] R. Zhang and Y. Wang, “An enhanced agglomerative fuzzy k-means clustering method with mapreduce implementation on hadoop platform,” IEEE International Conference on Progress in Informatics and Computing, 2016.

[20] Z. Qingyuan, Zhou. Zongming and W. Yuancong, “Wit120 data mining technology based on internet of things,” Health Care Management Science volume, pp. 680–688, 2020.

[21] Y. Zhong and D. Liu., “The application of k-means clustering algorithm based on hadoop.” IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBD&A), 2016.

[22] Y. K. Kim, Wooyeol and K. Shim., “Parallel computation of k-nearest neighbor joins using mapreduce.” International Conference on Big Data (Big Data), 2016.

[23] K. Sharmila and T. Kamalakannan, “Analytics for healthcare using hadoop map reduce, apache spark and in cloud services,” International Journal of Scientific & Technology Research, p. 706–710, 2020.

[24] S. Bethu, “Map reduce by k-nearest neighbor joins.” 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2018.

[25] Doreswamy, “Scalable k-means algorithm using mapreduce technique for clustering big data,” International Journal of Latest Trends in Engineering and Technology, Special Issue SACAIM 2016, 2016, p. 408–414, 2016.

[26] D. Chrimes, “Operational efficiencies and simulated performance of big data analytics platform over billions of patient records of hospital system,” Advances in Science, Technology and Engineering Systems Journal, pp. 23–41, 2017.

[27] X. Zhang and Y. Wang, “Research on intelligent medical big data system based on hadoop and blockchain,” EURASIP Journal on Wireless Communications and Networking, pp. 1–21, 2021.

[28] M. P. Prasanna Balaji Narasingapuram, “Sentiment classification of tourism based on rules and lda topic model.” International Journal of Recent Technology and Engineering (IJRTE), pp. 2277–3878, 2020.

[29] R. S. Sharon and R. J. Manoj, “E-health care data sharing into the cloud based on deduplication and file hierarchical encryption,” 2017 International Conference on Information Communication and Embedded Systems, 2017.

[30] H. Zhao, “Feasibility study on security deduplication of medical cloud privacy data,” EURASIP Journal on Wireless Communications and Networking, pp. 1–15, 2018.

USHARANI A V works for Applied Cognition Systems in Manipal as a software engineer. She graduated from VTU Belgaum with a bachelor’s degree in computer science and engineering. She worked as a software engineer for a few years before enrolling at MIT Manipal to pursue a master’s degree in information and communication technology in 2021.

GIRIJA ATTIGERI She is currently Assistant Professor-Selection Grade in the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. She has received B.E. and M. Tech. degrees from the Visvesvaraya Technological University, Karnataka, India. She has 15 years of experience in teaching and research. She has received his Ph.D. from the Manipal Institute of Technology, Karnataka, India. His research interests span big data analytics, Machine Learning and Data Science. She has around 10 publications in reputed international conferences and journals. She has supervised several UG and PG students.