A recommendation system based on AI for storing Block data in the Electronic Health Repository

Vinodhini Mani[1*], Kavitha c[1], Shahab S. Band[2*], Amir Mosavi[3*], Paul Hollins[4], selvashankar palanisamy[5]

[1]Computer Science Engineering, Sathyabama Institute of Science and Technology, India, [2]National Yunlin University of Science and Technology, Taiwan, [3]Óbuda University, Hungary, [4]University of Bolton, United Kingdom, [5]Ford Motor, India

Submitted to Journal:
Frontiers in Public Health

Specialty Section:
Digital Public Health

Article type:
Original Research Article

Manuscript ID:
831404

Received on:
08 Dec 2021

Revised on:
13 Dec 2021

Journal website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Author contribution statement

V.M and C.K: Conceptualization. V.M and C.K: Methodology, investigation, data curation, and writing—original draft preparation. S.S.B, M.A, H.P,P.S: software. S.S.B, M.A, H.P,P.S: validation and visualization. V.M M.A, H.P: formal analysis. S.S.B, M.A, H.P: resources. V.M, S.S.B, M.A: writing—review and editing, supervision. V.M and C.K, S.S.B, M.A, H.P: project administration. All authors have read and agreed to the published version of the manuscript.

Keywords

artificial intelligence, machine learning, Health Repository, Patients, Health data, storage, deep learning

Abstract

Word count: 151

A proliferation of wearable sensors that record physiological signals has resulted in an exponential growth of data on digital health. To select the appropriate repository for the increasing amount of collected data, intelligent procedures are becoming increasingly necessary. However, allocating storage space is a nuanced process. Generally, patients have some input in choosing which repository to use, although they are not always responsible for this decision. Patients are likely to have idiosyncratic storage preferences based on their unique circumstances. The purpose of the current study is to develop a new predictive model of health data storage to meet the needs of patients while ensuring rapid storage decisions, even when data is streaming from wearable devices. To create the machine learning classifier, we used a training set synthesized from small samples of experts who exhibited correlations between health data and storage features. The results confirm the validity of the machine learning methodology.

Contribution to the field

Patient health data privacy is an emerging area of interest nowadays. Although there are many blockchain-based storage systems available, the cost of storing data is prohibitive. Hence we have implemented a recommendation system for helping the patient to store their data based on user preference and doctor preference. This system has considered five machine learning algorithms and their performance is evaluated. Thus, our system strongly supports patients in storing their health data in health repositories.

Ethics statements

Studies involving animal subjects

Generated Statement: No animal studies are presented in this manuscript.

Studies involving human subjects

Generated Statement: No human studies are presented in this manuscript.

Inclusion of identifiable human data

Generated Statement: No potentially identifiable human images or data is presented in this study.

Data availability statement

Generated Statement: The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

A recommendation system based on AI for storing Block data in the Electronic Health Repository

Vinodhini Mani[1*], Kavitha.C[1], Shahab S. Band[2*], Amir Mosavi[3*], Paul Hollins[4], Selvashankar Palanisamy[5]

[1]Computer Science Engineering, Sathyabama Institute of Science and Technology, Tamilnadu, India

[2]Future Technology Research Center, College of Future, National Yunlin University of Science and Technology, 123 University Road, Yunlin 64002, Taiwan

[3]Institute of Software Design and Development, Obuda University, 1034 Budapest, Hungary

[4]Professor of Cultural Research Development School of Arts/ Institute of Management University of Bolton, Uk BL1 1SW

[5]Manager-Intelligent Automation, Ford Motors Pvt Ltd, India

*** Correspondence:**
Corresponding Author
vinodhini.cse@sathyabama.ac.in, shamshirbands@yuntech.edu.tw, amir.mosavi@nik.uni-obuda.hu

Keywords: Artificial Intelligence, Deep Learning, Health Repository, Health data, Machine Learning, Storage

Abstract

The proliferation of wearable sensors that record physiological signals has resulted in an exponential growth of data on digital health. To select the appropriate repository for the increasing amount of collected data, intelligent procedures are becoming increasingly necessary. However, allocating storage space is a nuanced process. Generally, patients have some input in choosing which repository to use, although they are not always responsible for this decision. Patients are likely to have idiosyncratic storage preferences based on their unique circumstances. The purpose of the current study is to develop a new predictive model of health data storage to meet the needs of patients while ensuring rapid storage decisions, even when data is streaming from wearable devices. To create the machine learning classifier, we used a training set synthesized from small samples of experts who exhibited correlations between health data and storage features. The results confirm the validity of the machine learning methodology.

1 Introduction

In the modern era, clinicians no longer manage health data exclusively, but are increasingly responsible for obtaining consent from patients (1). The rights of patient's access to, analysis of, and exchange of their health information have evolved dramatically (2). The majority of patients are dissatisfied with their health care providers after sharing self-tracking data (3). It is still possible to enhance patient health care by incorporating patient health data into the current health data systems. Literature has identified various categories of patient health information (4). These categories include information about medications, biometrics, behavioral information, data about social interactions, genetics, psychological data, data about symptoms, and reports. Blockchain-based interplanetary file system secondary storage of health data has been implemented to safeguard the privacy and security of patient health information (5).Yet very few studies have evaluated how patients' health data is stored. A key component of the proper management of health data is protecting the privacy and confidentiality of the patient while maintaining data accessibility for relevant stakeholders. Studies indicate that health data security poses a massive threat. This is evidenced by the proliferation of medical devices with limited memory and power (6, 7) and substantial medical data repositories (8).Many types of organizations are responsible for managing the massive amount of health data.

Health data is often portrayed as being sensitive to all patients with the same level of privacy and confidentiality; however, this is not true in practice because it is not equally sensitive to everyone at the same time. When a patient reaches a high level of public prominence, she may surrender the ECG data she generated on her own and to her cardiologist. This data can be accessed by other healthcare providers through an electronic health record. A patient who wishes to keep her pregnancy test results private may be forced to allow her provider to store her pregnancy test results. The dissemination of health data between multiple providers who manage data repositories now enables the storage medium to be customized based on patient needs. This includes the cost, size, security, confidentiality, and privacy of each chunk of data. Hybrid execution models, such as those described by the author (9), allow sensitive data to be stored in private clouds while no sensitive data is maintained in public clouds. Nevertheless, it does not specifically address health data processing. Communication between the two cloud platforms also takes time, and computations that rely on bandwidth use a lot of resources. A hybrid cloud platform was developed by (10) for solving this problem. Medical sensors, apps, and devices provide data to artificial intelligence, which enables the automatic diagnosis of health conditions. Health data, including ECG, blood pressure, and pulse rate, can be classified as normal or abnormal by algorithms based on a range of conditions and thresholds set by healthcare professionals. Clinical research and clinical care are usually aided by abnormal data. Using the Body Area Sensor Network, (8) developed an agent-based system developed for elderly people to preserve abnormal data. Health information is generated in enormous quantities nowadays, so a diverse storage solution is needed(11) .Several researchers have examined the performance and cost parameters of various Cloud Service Providers (CSPs) to design methods for selecting suitable CSPs for storing consumers' data(12,13,14) .High-performance cloud services minimize the time spent in operations but incur high costs. Additionally, researchers are investigating blockchain technology for its promise of security and privacy for health data management. Combining blockchain-based eHealth with traditional health databases is possible, which can be arranged based on users' preferences and the possibility of utilizing the data in the future. However, due to the design of blockchains, they are not suitable for hosting large amounts of health data. A software agent that knows the patient's preferences is inserted inside the application in (15).Nonetheless, they never described a way to make this decision. To assist in choosing storage repositories, we developed a model that incorporated not only (8)'s criteria, but also aspects like data confidentiality, privacy, and quality of performance.

Motivation

Every Blockchain miner owns a local ledger, so this technology allows transactions to be verified and processed without the need for third parties. Verifying transactions does not require a centralized server. Document alterations cannot be guaranteed through conventional database storage and blockchain-based hash management. Data is only detectable in a blockchain if a hash pointer holds a pointer to it. Depending on the patient, personal preferences, and other factors, the sensitivity and significance of the health information are also different from repository to repository. Choosing the right repository is extremely crucial. As wearable sensors continuously stream health data, the challenges are exacerbated. In (16), the author has surveyed the importance of artificial intelligence in healthcare. The prediction of COVID-19 infected patients using artificial intelligence has been implemented in (17), but there is a need for an appropriate repository to store the data.

Contribution

In our research, we considered the variation in data sensitivity, volume, and other factors to locate the appropriate system to manage health records. The flow diagram of the paper contribution is shown in Figure 1. Collect the health data and health repository parameters. Evaluations of both health information and health repository parameters are given a score. The machine learning-based recommendation model for health data storage proposes a way to distribute health data among multiple repositories. A model for automated health data storage recommendation is being developed to determine appropriate storage repositories. Through correlation analysis, user preferences, and clinical heuristics, a machine learning-based classifier is used to map health data characteristics to each repository. Patients' security and privacy preferences are taken into account as well as the sensitivity of health data.

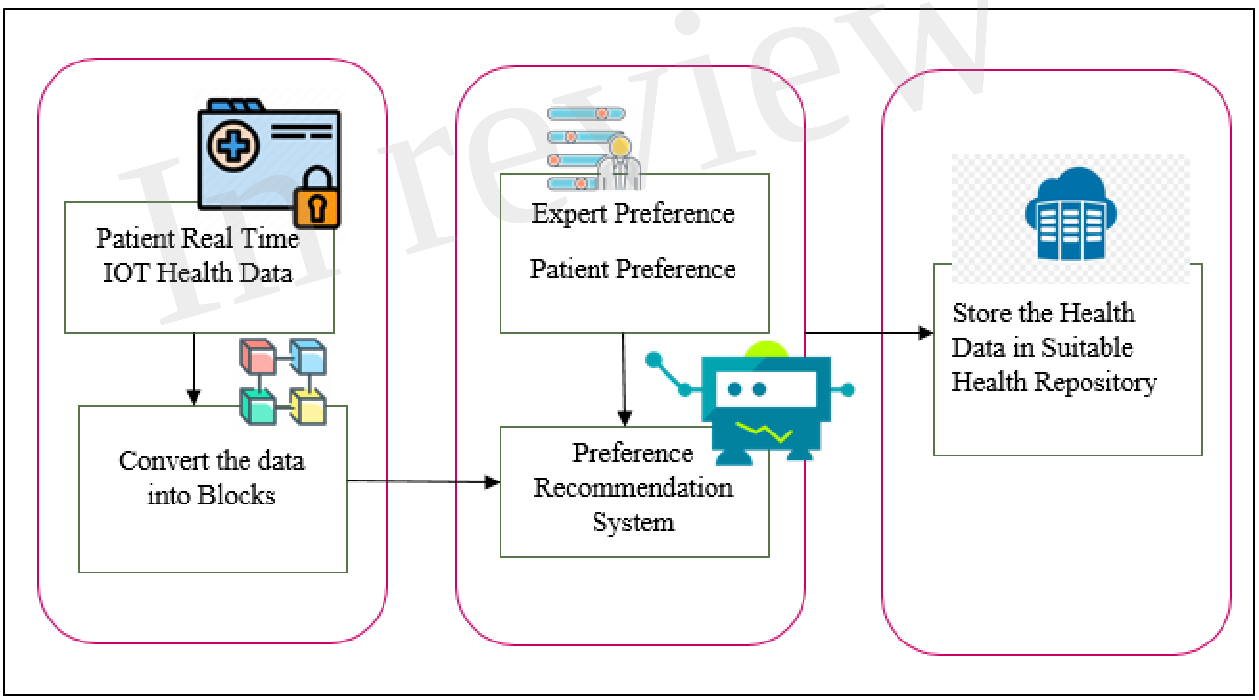


Figure 1 Paper Contribution Flow Diagram

Organization

Following are the sections of the paper: Section 2 addresses related work. In Section 3, we present the proposal for a recommendation model for a health repository. Section 4 describes how the system will be implemented. The results and evaluation of performance will be discussed in Section 5. Conclusions and future work will be discussed in Section 6.

2 Background

Big Data cannot be stored, accessed, or analyzed with a single health record system. Patients can lose medical information when their electronic health records are malfunctioning (18).Due to the manual uploading of data generated by wearable sensors to personal health records, caregiver responses were delayed. For this reason, (19) developed methods for storing patient-generated health information on commercial blood glucose monitors. The electronic health record system could be made to fit the streamed data if it is filtered or compressed (20).In (21, 22, 23, 24), a number of action plans and standards were advocated for the adoption of an electronic health record system. A selection of an electronic health record should take into account functional requirements, troubleshooting, and optimization features (22).The author provides a list of steps to follow before buying an electronic health record system. Checklists mostly cover client meetings on site, site visits, and maintaining live workflows. Health data sources such as hospitals, clinics, insurers, and patients should be integrated into centralized databases, according to the author (25). In particular, patient-centered health data with high degrees of structural heterogeneity must be stored and processed quickly because of their high volume and rate. For health data, to provide useful insights, precision is essential, but some sources produce vague and inaccurate information. Distributed data storage systems do offer some relief to these issues. (26)Various cloud storage mediums have been examined. A machine learning and deep learning model is used to predict the thermal sensation vote system (27). Utilization of a compression algorithm to retrieve the health repository data as fast as possible using blockchain and interplanetary file systems (IPFS) without data loss (28). Diabetic Retinopathy is efficiently classified using a deep learning and machine learning algorithm (29). Genetic algorithm with fuzzy logic is a tool to help medical practitioners diagnose heart disease at an early stage using adaptive genetic algorithm with fuzzy logic (AGAFL)(30) .Health data storage systems and data properties were not considered in the selection of repositories. Furthermore, no machine learning mechanisms were developed to cater to user preferences.

In the next section, we describe how we facilitate distributed health data management.

3 Model for Recommendation of Health Repositories

As data streams increase, the need for storage decisions becomes more frequent, making manual consultation with patients an inefficient process that requires an automated solution. It is, however, impossible to prespecify the data storage requirements for each patient that will apply to all possible future contexts. The learning classifier may generalize to a broader range of mappings based on a manual mapping specification by an expert.

The following sections explain in detail the overall approach described in Figures 2 and 3. Data storage requirements - an illustration of which is displayed in layer 1 of Figure2, consists of a set of variables or features that characterize the requirements for storing a chunk of data. Some of the attributes' values have been shown to be numerical (1 - 10) and others to be qualitative. Secondly, each instance of the dataset contains the specifications required to store each chunk of data as shown in Figure 2.

Health Repository Evaluation Criteria are calculated in layer 3 by adding a rating provided by an expert group. These criteria reflect the characteristics of storage repositories as shown in Figure 2. Three standards apply to rank five storage repositories. Medical professionals and patients themselves may create clinical heuristic rules in layer-3 of Figure 2 and each instance in the dataset is categorized according to the preferences of the users. A storage repository can be assigned to an instance based on heuristic rules in a real-world situation. The correlation coefficient offers an inference of a class label when preferences and heuristics do not match well. The health repository requirements can be mapped to layer-4 (user and expert expectations) by a machine learning classifier, as shown in Figure 2.In

Figure 3, a recommendation framework for health repositories is illustrated. There are two parts to the framework: determining which standards should be used for the storage and assessment of data and implementing machine learning.

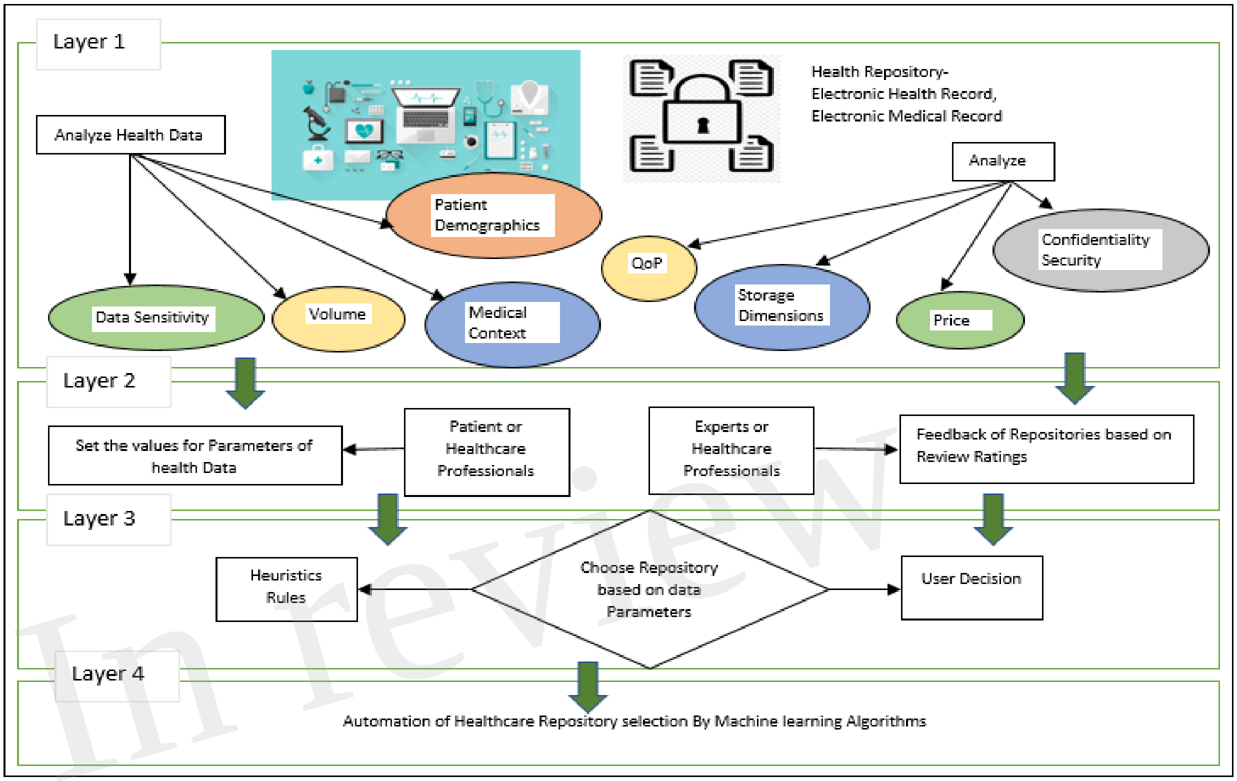


Figure 2 Proposed System Architecture

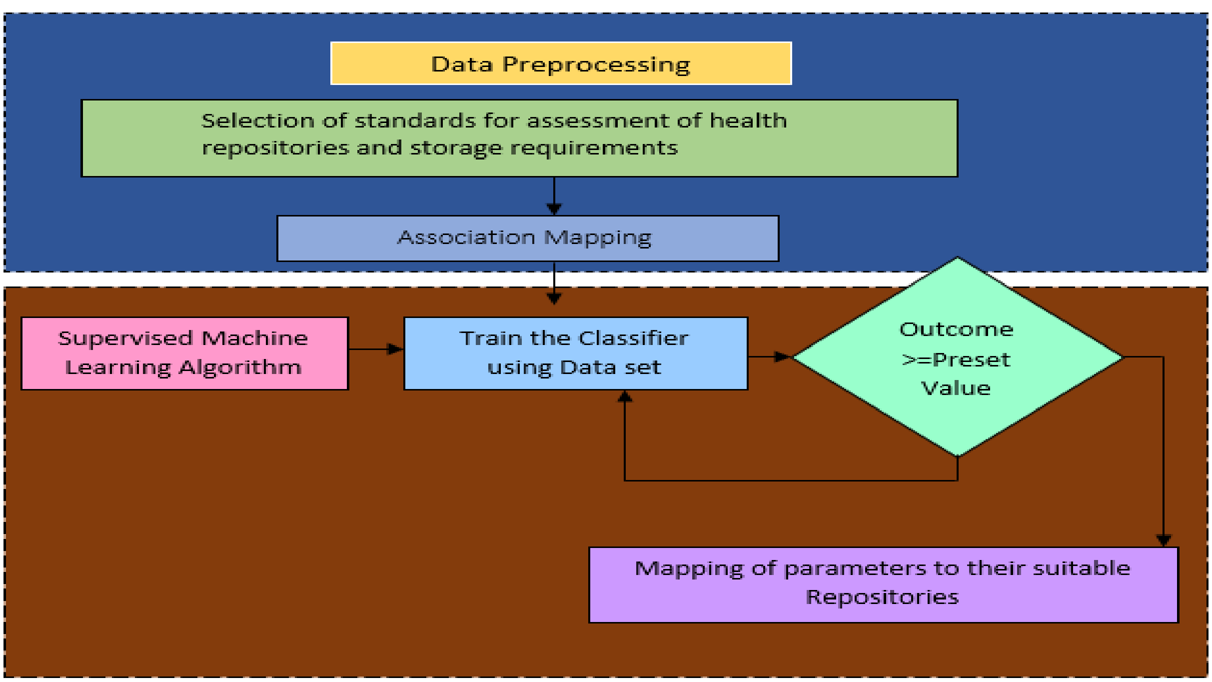


Figure 3 Proposed Health Repository Recommendation System

4 Implementation

This recommendation system assumes that a patient is in full control of his or her decision regarding storage. It is impossible to make decisions manually in many cases because they are made so frequently. Hence, automated processes are essential. In the mapping process, the characteristics of a repository managed by an agent group are matched with the characteristics of data about the storage requirements of patients. Because patients' storage requirements vary so much, it is impossible to predetermine every possible scenario. By utilizing a set of mappings that is specified manually by experts, machine learning is used to generalize a mapping over a wide range of patient contexts. This methodology involves defining a set of attributes that describe what chunk of data needs to be stored. There are numerical values and categorical values assigned to those attributes. Thus, a dataset containing these attributes will be created, with each instance representing a different set of storage requirements. A group of experts' ratings are then used to determine the characteristics of the available storage mediums. To determine what class each instance falls into, statistical correlation and heuristic rules are employed. Based on the training datasets, the supervised machine learning classifier maps the data into a storage repository. Figure 3 illustrates two components of the recommendation system: Data Pre-processing and Supervised Machine Learning. According to Figure 3, the upper portion of the framework contains the characteristics of the data storage requirements. There are a number of features that demonstrate the characteristics of health repositories. A number of associations were found between the two groups of features.

4.1 Data Preprocessing

The data collected from hospitals and patients undergoes a preprocessing process, which includes analyzing data storage requirements, identifying sensitive data areas, analyzing the volume of each record, analyzing the patient health profile, determining the demographics of patients, and analyzing health repository parameters as well as storage, cost, security, privacy, and performance.

4.1.1 Characteristics of data storage requirements

To determine which repository is the best option, consideration is given to the sensitivity of the data, the volume of the data, medical care context, and demographics of the patient.

4.1.1.1 Sensitivity of the data

It is imperative to prevent unauthorized access to all health-related data. Depending on the data type, some breaches are more likely than others. Depending on the individual's preferences and context, the level of data sensitivity may vary.

4.1.1.2 The volume of the data

Reports, medical diagnoses, and medication summaries are not frequently created, which means that their storage needs are less than those of health data sets.

4.1.1.3 Context of Medical Care

The context may be palliative care, critical care, chronic illness, or no chronic illness. The context may also differ based on the country.

4.1.1.4 Demographics of patients

Several factors can play a significant role in determining which storage medium to use, such as socioeconomic status, occupation, education, and nationality.

4.1.2 Health Repository Evaluation Parameters

Evaluation parameters for health repository such as security, privacy, cost, storage capacity, and performance. Table 1 shows the parameters and criteria of the health repository evaluation.

Table 1 Health Repository Evaluation

Assessment Parameters	Survey Questions for Health Repository Ratings
Storage	Can the repository be used to store Big Data?
	Regarding processing Big Data, what is the repository's role?
	Are there any benefits to storing continuously streamed data in the repository?
Cost	Does deployment cost a lot?
	Does maintenance cost much?
	What is the service cost?
Security	Is the storage repository capable of maintaining data integrity?
	Does the storage repository have 24/7 accessibility?
	Are storage repositories resistant to cyberattacks?
Privacy	Is data accessible to third parties?
	Is the access control right given to the owner of the health records?
Performance	How fast can you upload files?
	Is it possible to retrieve data quickly?
	Is it possible to process data quickly?

4.1.3 The relationship between repository evaluation standards and data features

Medical records, in particular those generated by patients, are to be transferred to a health record system that reflects the preferences of the user and the data requirements. Health data requirements and criteria for evaluating storage are correlated in a one-to--to-many fashion as implemented in Algorithm 1.Some associations are strong, and some are weak. To facilitate the rapid processing of highly confidential data, a health record system may accept data blocks in plaintext format. Data with relatively low confidentiality can be highly sensitive due to the demographic characteristics of patients. Data about a patient's demographics, such as their educational background and professional experience, may affect their privacy concerns. Users can then choose from a variety of storage repositories that protect their confidentiality. The sample association mapping as shown in Table 3.

Table 3 Association Mapping

S.No	Characteristics of data storage requirements	S.No	Health Repository Evaluation Parameters	Association Mapping
1	Sensitivity of the data	A	Storage	1→(B,C,D,E)
2	The volume of the data	B	Cost	2→(A)
3	Context of Medical Care	C	Security	3→(E)
4	Demographics of patients	D	Privacy	4→(B,C,D,E)
		E	Performance	

4.2 Supervised Machine Learning Algorithm

Dynamically suggest health repositories based on supervised learning for particular data blocks, which is implemented using Algorithm 2. A training dataset must be generated for every instance of the dataset in addition to the labeled training datasets. Health repositories will be assigned data blocks that have a number of attributes. Among the attributes are some that are directly linked to the data block and others that are directly linked to the patient. Attributes include data sensitivity, volume, context of care, and demographics of the patients. The health repository should consider for evaluation such as electronic health records, cloud based electronic health records, blockchain based electronic health records, patient health record, and Electronic Medical Records. We considered the following health repository parameters in this study: security, privacy, cost, storage capacity, and performance. Each repository has been assigned a rating value ranging from 1 to 10.Whenever other attributes are not significant in determining the health repository, a linear regression Y (15) is calculated to label the instance as shown in Equation 1.

$$Y=A+RX \qquad (1)$$

$$R=n\left(\sum_{i=1}^{n} xiyi - \left(\sum_{i=1}^{n} xi\right)\left(\sum_{i=1}^{n} yi\right)\right) \qquad (2)$$

$$A=\frac{\left(\sum_{i=1}^{n} yi\right)-R\left(\sum_{i=1}^{n} xi\right)}{n} \qquad (3)$$

Where R is the Coefficient which contains R1,R2,R3,R4,R5,R6,R7,R8,R9,R10 are calculated between the set of data storage requirements(DR) as shown in equation 2. Here are the evaluation criteria for Electronic health record (D1), Patient health record (D2), Cloud-based electronic health record (D3), Blockchain-based electronic health records (D4), and Electronic Medical records (D5). The calculation of health repository recommendation Di is estimated using the equation

$$Di=High\ (R1, R2.....Rm) \qquad (4)$$

M is the number of health repositories and n is the rating criteria. Secondly, the choice of a health data repository can be influenced by the decision of the healthcare professional, the preferences of the user, and a variety of factors such as normal or abnormal behavior patterns and patient health status, as well as other demographic factors. Patients with unusual health patterns should store their health records in a repository that health care professionals can access quickly. A less secured and less expensive repository can be used to store data which is hardly ever accessed by health care professionals. Different users may have different privacy preferences, and those preferences may change over time based on different contexts (31). The health record system for a patient should take into account a variety of factors. There are several factors involved, such as medical conditions, personal characteristics, socioeconomic status, as well as the type and significance of data. The level of privacy and security preferences of individuals may change over time as well. In contrast to patients with terminal illnesses, young individuals may be more concerned with privacy and security. By considering author preference, some of the sample user preference and health professional preference heuristic rules were implemented, as shown below

1. If (Data= standard && volume=large)

Then

Storage Repository=Cloud based Health Record Management System

2. If (Data= standard && volume=low)

Then

Storage Repository=Blockchain enabled Personal Health Record System

3. If (Data=Unusual patterns && volume=low)

Then

Storage Repository=Blockchain based Electronic Medical Record

4. If (Patient= Famous Personality && health condition = Good))

Then

Storage Repository=Blockchain based Electronic Health Record

5. If (Patient= Famous Personality && health condition = Serious))

Then

Storage Repository=Blockchain based Electronic Medical Record

6. If (Data of type Disease)

Then

Store data in Disease Registry

<u>Algorithm 1: Association mapping ()</u>

```
Step 1:   Begin
Step 2:   Let Data Source as DS;
Step 3:   Let Storage Requirements as SR;
Step 4:   Let Health Repository Parameters as HRP;
Step 5:   For each data ϵ DS do
Step 6:      For each Storage Requirement ϵ SR do
Step 7:          Collect the data;
Step 8:          Identify the SR;
Step 9:          Collect the HRP;
Step 10:     For each SR and HRP do
Step 11:         Analyze the parameters using Evaluation Criteria;
Step 12:         If (SR ϵ HRP)
Step 13:               SR (SR1…n)→ HRP (HRP1…n);
Step 14:                Create Association Dataset as AD;
Step 15:         Else
Step 16:                 Print Not Associated;
Step 17:         End; End; End; End; End;
```

Algorithm 2: Health repository Recommendation system ()


```
Step 1:   Begin
Step 2:   data collected from various data sources;
Step 3:   Call Association Mapping ();
Step 4:   For each Health Data Block ϵ HB do
Step 5:       Select the Supervised Machine learning algorithm;
Step 6:       Train the Data block HB;
Step 7:       Apply Heuristic Rule;
Step 8:       If (Accuracy >= Threshold)
Step 9:           Test data;
Step 10:          Allocate the Health Data Block HB → Health Repository HR;
Step 11:          Send (Recommend Repository to Patients);
Step 12:          Break;
Step 13:      Else
Step 14:          Continue;
Step 15:     End; End; End;
```

5 Results and Discussion

Research was conducted on supervised machine learning classification techniques. Using the WEKA tool, different classification algorithms were tested. The study used an Intel Core i7 6700H processor with up to 3.5 GHz and 16 GB of RAM. The dataset was divided into training and test sets. Data preprocessing is performed prior to analysis. To train the data in the recommended health repository, linear regression data blocks and user and health professional preference rules have been used. During this experiment, we determine whether the classifiers can learn how to classify data distributions. The training datasets each contain 400, 800, 1200, and 2000 instances. Table 3 shows the mapped sample training dataset.

Four different classifiers were run on four datasets to test whether a machine learning algorithm could choose an appropriate storage medium, NaïveBayesSimple, Multilayer Perceptron, Random Forest

Classifier, Random Tree and the IB1 algorithm are four different types of classifiers trained here. Several classification techniques were compared using Python to determine their accuracy scores (32).

Table 3 maps Sample Training Data set

Information block	**Sensitivity data**	**Volume**	**Context of Medical Care**	**Social Status**	**Profile Visibility**	**Patient Status**	**Health Repository**
Data Block1	1	2	3	3	high	Typical	Blockchain based Electronic Health Record
Data Block2	2	5	3	5	Low	Typical	Cloud Electronic Health Record
…….	…	…	…	….	…	….	…..
Data Block n	3	2	3	2	1	Abnormal	Electronic Medical Record

5.1 Classification Model accuracy

1. Confusion Matrix

2. Classification Measure

5.1.1 Confusion Matrix

In the confusion matrix, N is the number of target classes, and N is the number of rows. It is used to evaluate the performance of a classification model. Machine learning is used to predict target values from the actual values in the matrix. True Positive (TP) and True Negative (TN) rates should be high and False Positive (FP) and False Negative (FN) rates are low for a successful model. A confusion matrix as is always more appropriate as a machine learning model evaluation criterion when working with an imbalanced dataset.

5.1.2 Classification Measure

As an evaluation measure, the classification measure is used in addition to the confusion matrix. They are

1. Accuracy

$$\textbf{Accuracy} = \frac{\textbf{TP+TN}}{\textbf{TP+TN+FP+FN}} \quad \mathbf{0.0 < Accuracy < 1.0} \qquad \textbf{(5)}$$

2. Precision

$$\textbf{Precision}=\frac{\textbf{TP}}{\textbf{TP+FP}} \qquad \textbf{(6)}$$

3. Recall

$$\textbf{Recall}=\frac{\textbf{TP}}{\textbf{TP+FN}} \qquad \textbf{(7)}$$

4. F1-Score

$$\textbf{F1-Scrore=2}\,\frac{\textbf{Recall}*\textbf{Precision}}{\textbf{Recall+ Precision}} \qquad \textbf{(8)}$$

5. Sensitivity and Specificity

$$\textbf{Sensitivity}=\frac{\textbf{TP}}{\textbf{TP+FN}} \qquad \textbf{(9)}$$

$$\textbf{Specificity}=\frac{\textbf{TN}}{\textbf{TN+FP}} \qquad \textbf{(10)}$$

6. Root Mean Square Error

Modified Mean Square Error (MSE) is a variation of Root Mean Square Error (RMSE). Measuring the mean square error squared is equivalent to this metric. The RMSE of an ideal model is zero, just as the MSE and MAE are zero.

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\textbf{Actual Values} - \textbf{Predicted Values})^{2}} \qquad (11)$$

5.1.3 Result Analysis

As illustrated by the graph in Figure 4, Random Forest classifiers become more accurate as the number of instances increases, as shown by a 10-fold cross-validation analysis. A balanced ratio of each class was found in the dataset of 1200 records, thus all classifiers performed better. The Random Forest performed best, with 98.21% accuracy. On the 2000-record dataset, however, all classifiers had lower accuracy, largely because the dataset was skewed. Compared to other classifiers, Random Forest exhibits lower root mean square error in Figure 5. Figure 6 illustrates the percentage split results, which are less accurate than the cross-validation results presented in 10-fold cross-validation.

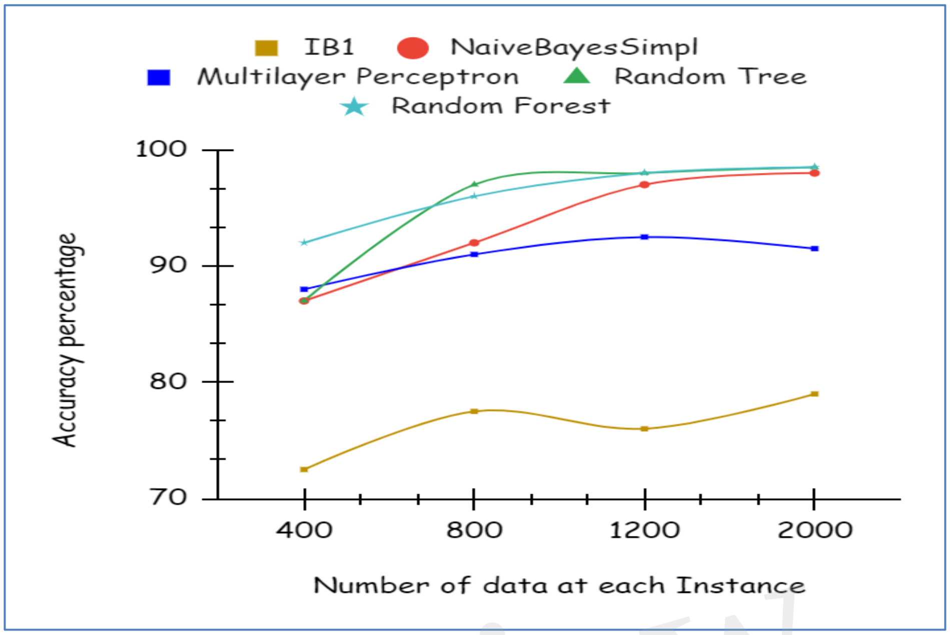


Figure 4 Accuracy Using 10-Fold cross validation

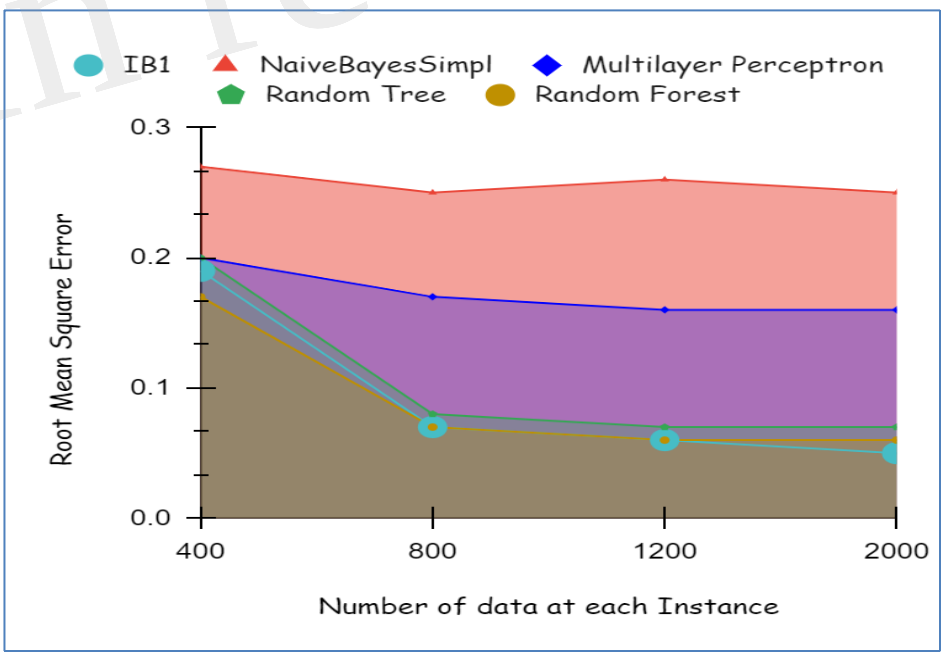


Figure 5 RMSE Using 10-Fold cross validation

By using a percentage split, 80% of the data were used for training and 20% for testing. The classifier is trained only once, as seen in Figure 7, which demonstrates low accuracy and large RMSE. Artificial intelligence is a technique for deep learning.

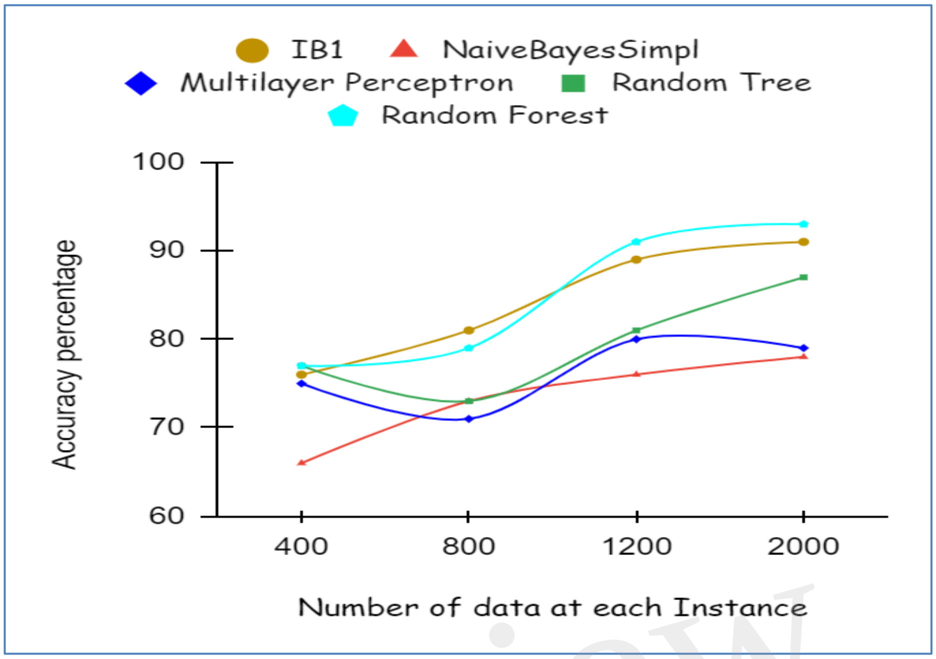


Figure 6, Accuracy of Percentage split dataset

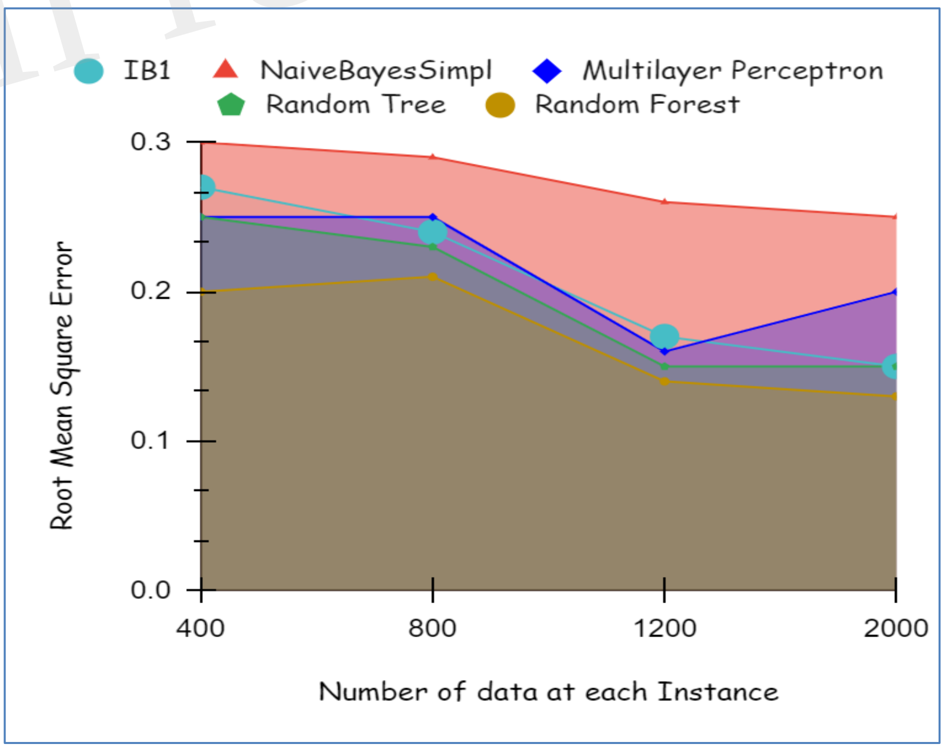


Figure 7 Accuracy of Percentage split dataset

Using deep learning networks, unstructured or unlabeled data can be learned unsupervised. Real-world health repositories are usually recommended based on unstructured and unlabeled datasets. For our synthetic dataset, we analyzed the accuracy using a deep learning algorithm. A deep learning model is run on the synthetic dataset, and it shows 88.70 percent accuracy. It is implemented in Python. There are three hidden layers in the model; the first of these layers has 100 output nodes, while the second and third have five output nodes each. Training is done with 100 iterations and eight batches are used. The training dataset is shown in Figures 8 and 9, with a Y-axis showing the loss and X-axis showing the number of iterations. A deep learning classifier and a machine learning classifier are displayed in Figures 10, 11, and 12 for the classification. With reference to recall, F1-measure, and precision, the Random Forest classifier outperformed the other tested classifiers. Classes that were allowed and those that were not were included in the experiment. In terms of recall, precision, and F1-measure, the Random classifier scored 93, 100, and 96% for cloud electronic health records, 100, 92, and 96 for blockchain-based electronic health records, and 85, 96, and 90 for electronic medical records. In terms of the allowed class, the rest of the experimented models perform well. In terms of the disallowed class, they did not perform well.

Figure 8 Performance loss of Training and Test Set

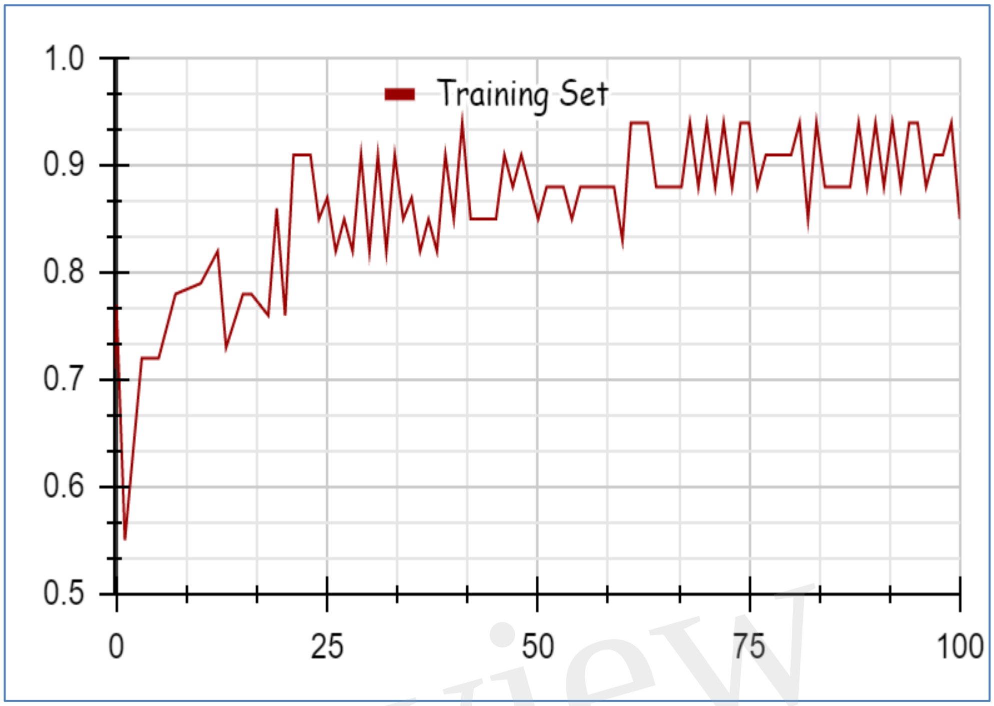


Figure 9 Performance Accuracy of Training and Test Set

The accuracy of the classifier supports the use of machine learning to map the health storage mediums to health data blocks. Given the growing volume of health data that will need to be stored and accessed globally, this machine learning model may play a crucial role in improving storage and access arrangements in the future. This will make health data storage easy and straightforward for consumers. In addition, they would be able to ensure that the size of the data store is manageable. It can help to determine which storage solution best fits the requirements of different data assets using a machine learning model.

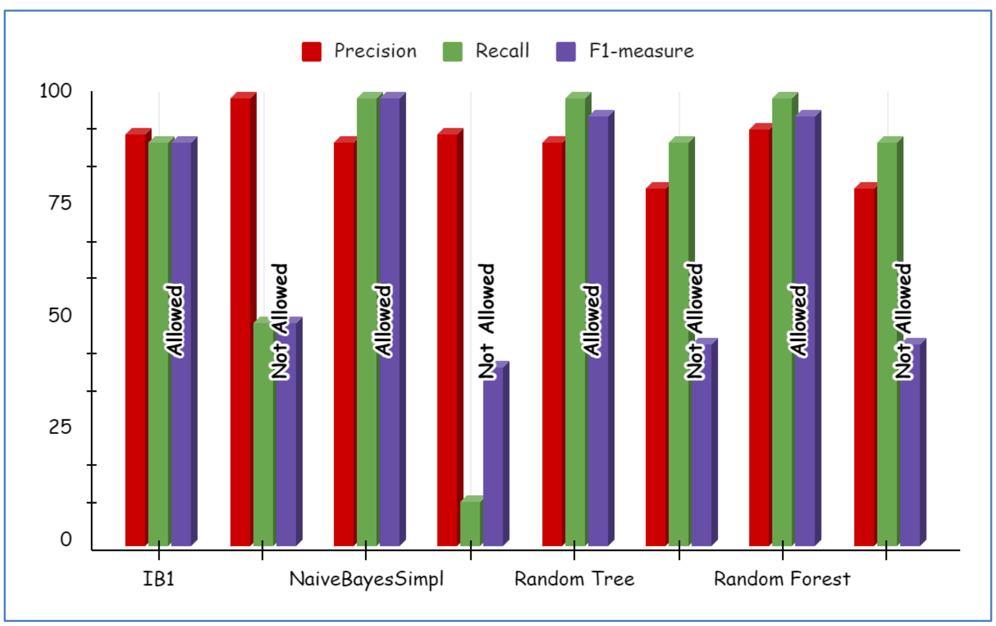


Figure 10 Deep learning results for Cloud Electronic Health Record

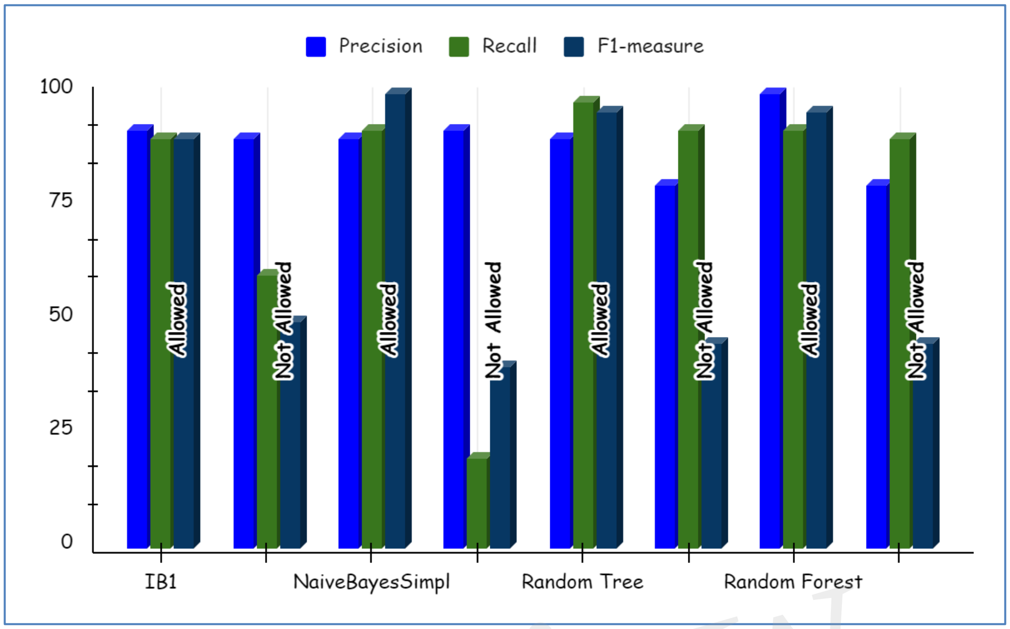


Figure 11 Deep learning results for Blockchain based Electronic Health Record

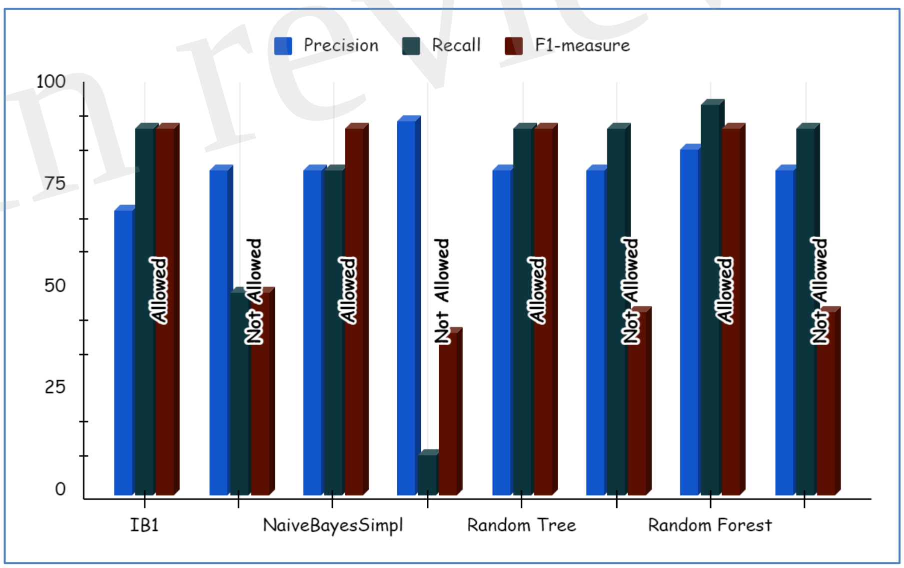


Figure 12 Deep learning results for Electronic Medical Record

5.1.4 Mapping of health data parameters to repositories

Medical technology is expected to develop health record systems in the future. Health records are taking on novel forms as a result of the expansion of medical data. As described below, the proposed system will support various data variations and health records. First, the system requests the ratings for the latest health record on the basis of health parameters from the IT staff and healthcare professionals. Second, the system relabels instances from the entire training dataset. As soon as a new instance is created, the old instances' labels do not change.

6 Conclusion

Health data will increasingly be preserved in a variety of repositories, so patients can select the repository that best meets their needs. Patients are realistically expected to avoid using a single repository for all their health data because the context of treatment, patterns of data, and legal constraints may change. To automate the storage decision, a selection algorithm must be developed. This is especially relevant in the case of constantly streaming health data. The process of choosing the right repository is complicated. In addition to knowledge of storage features used for interoperability, data security, and privacy, regulatory concerns must also be considered. To preserve confidentiality, we propose distributing health data among various vendors. By keeping medical records together, confidentiality will also be preserved. Based on factors like data type, sensitivity level, significance, patient safety, and privacy requirements, this model can recommend which health data blocks should be stored on which storage medium. When applied to the dataset generated, random forest yielded the highest accuracy of 96.4%. Accuracy of algorithms depends on the dimension, origin, and nature of the data. As a result, we intend to evaluate these various algorithms with different characteristic datasets in the near future. In the future, we will implement a role-based access control system to store medical record information by integrating the health repository recommendation system to allow access to the health records based on the permission of patients.

7 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8 Author Contributions

V.M and C.K: Conceptualization. V.M and C.K: Methodology, investigation, data curation, and writing—original draft preparation. S.S.B, M.A, H.P: software. S.S.B, M.A, H.P: validation and visualization. V.M M.A, H.P: formal analysis. S.S.B, M.A, H.P: resources. V.M, S.S.B, M.A: writing—review and editing, supervision. V.M and C.K, S.S.B, M.A, H.P: project administration. All authors have read and agreed to the published version of the manuscript.

9 Funding

This is funded by College of Future, National Yunlin University of Science and Technology, 123 University Road, Yunlin 64002, Taiwan

10 Acknowledgments

We acknowledge College of Future, National Yunlin University of Science and Technology, 123 University Road, Yunlin 64002, Taiwan for supporting this study.

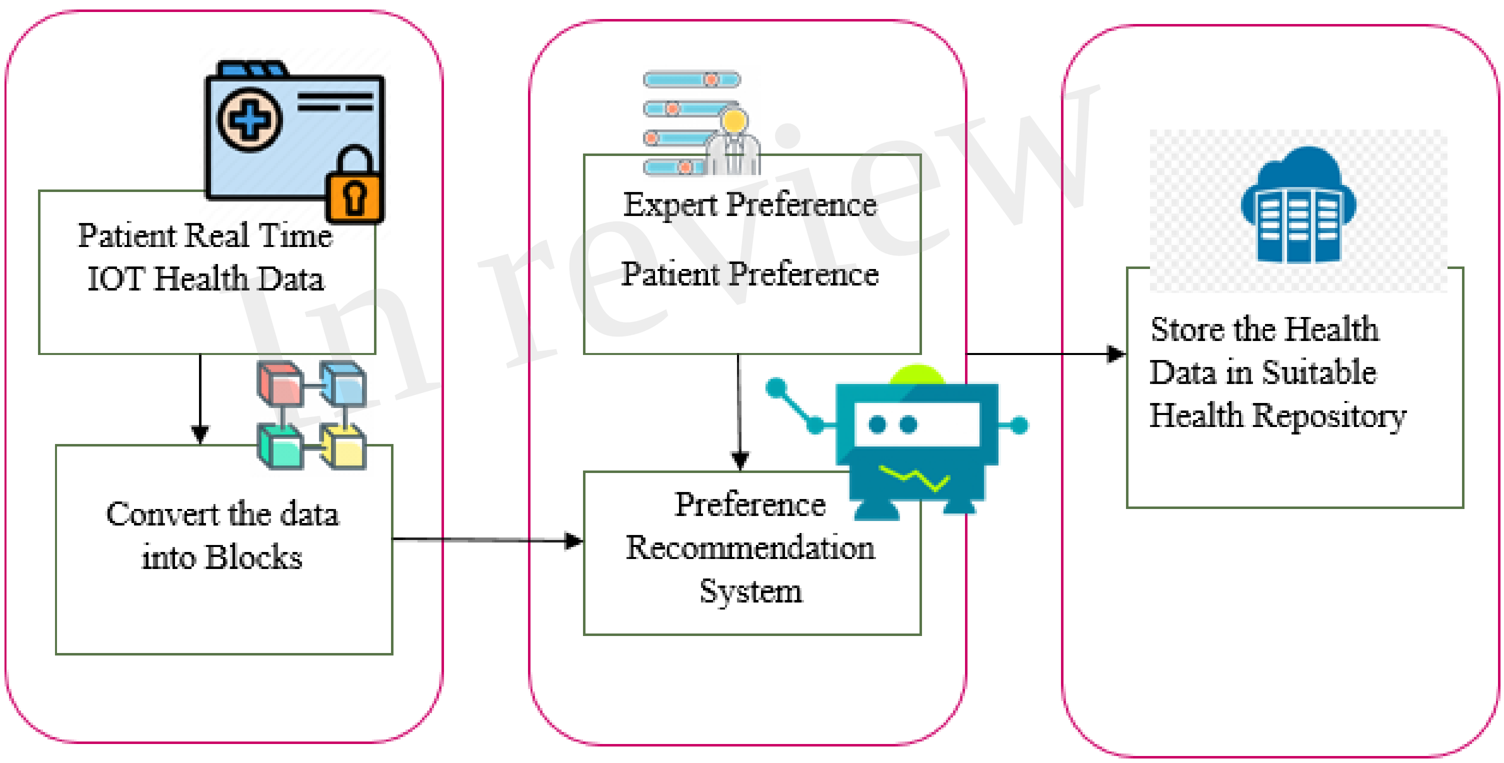
11 References

1. P. Plastiras and D. O'Sullivan, "Exchanging personal health data with electronic health records: A standardized information model for patient generated health data and observations of daily living," International journal of medical informatics, vol. 120, pp. 116–125, 2018.
2. A. Cortez, P. Hsii, E. Mitchell, V. Riehl, and P. Smith, "Conceptualizing a data infrastructure for the capture, use, and sharing of patient-generated health data in care delivery and research through 2024 (white paper)," 2018.
3. C.-F. Chung, K. Dew, A. Cole, J. Zia, J. Fogarty, J. A. Kientz, and S. A.Munson,"Boundary negotiating artifacts in personal informatics: Patient-provider collaboration with patient generated data," in Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 2016, pp. 770–786.

4. R. J. Lordon, S. P. Mikles, L. Kneale, H. L. Evans, S. A. Munson, U. Backonja, and W.B. Lober, “How patient-generated health data and patient-reported outcomes affect patient–clinician relationships: A systematic review,” Health Informatics Journal, p. 1460458220928184, 2020.
5. Mani, V., Manickam, P., Alotaibi, Y., Alghamdi, S., & Khalaf, O. I. (2021). Hyperledger Healthchain: Patient-Centric IPFS-Based Storage of Health Records. Electronics, 10(23), 3003.
6. A. Albahri, A. Zaidan, O. Albahri, B. Zaidan, and M. Alsalem, “Real-time fault-tolerant mhealth system: Comprehensive review of healthcare services, opens issues, challenges and methodological aspects,” Journal of medical systems, vol. 42, no. 8, p. 137, 2018.
7. D. Isern and A. Moreno, “A systematic literature review of agents applied in healthcare,”Journal of medical systems, vol. 40, no. 2, p. 43, 2016.
8. V. Vaidehi, M. Vardhini, H. Yogeshwaran, G. Inbasagar, R. Bhargavi, and C. S. Hemalatha, “Agent based health monitoring of elderly people in indoor environments using wireless sensor networks,” Procedia Computer Science, vol. 19, pp. 64–71, 2013.
9. S. Y. Ko, K. Jeon, and R. Morales, “The hybrex model for confidentiality and privacy in cloud computing.” HotCloud, vol. 11, pp. 8–8, 2011.
10. H. Zhang, L. Ye, X. Du, and M. Guizani, “Protecting private cloud located within public cloud,” in Global Communications Conference (GLOBECOM), 2013 IEEE. IEEE, 2013, pp. 677–681.
11. A. Stranieri and V. Balasubramanian, “Remote patient monitoring for healthcare: A big challenge for big data,” in Managerial Perspectives on Intelligent Big Data Analytics. IGI Global, 2019, pp. 163–179.
12. A. Ruiz-Alvarez and M. Humphrey, “A model and decision procedure for data storage in cloud computing,” in Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012). IEEE Computer Society, 2012, pp.572–579.
13. “Toward optimal resource provisioning for cloud mapreduce and hybrid cloud applications,” in Proceedings of the 2014 IEEE/ACM International Symposium on Big Data Computing. IEEE Computer Society, 2014, pp. 74–82.
14. M. S. Yoon and A. E. Kamal, “Optimal dataset allocation in distributed heterogeneous clouds,” in 2014 IEEE Globecom Workshops (GC Wkshps). IEEE, 2014, pp. 75–80. P. Plastiras and D. O’Sullivan, “Exchanging personal health data with electronic health records: A standardized information model for patient generated health data and observations of daily living,” International journal of medical informatics, vol. 120, pp. 116–125, 2018.
15. Zhang, Q., Lu, J. & Jin, Y. Artificial intelligence in recommender systems. Complex Intell. Syst. 7, 439–457 (2021). https://doi.org/10.1007/s40747-020-00212-w
16. Y. Yang and T. Chen, “Analysis and visualization implementation of medical big data resource sharing mechanism based on deep learning,” IEEE Access, vol. 7, pp. 156 077–156 088, 2019.
17. Stock C, Dias S, Dietrich T, Frahsa A and Keygnaert I (2021) Editorial: How can We Co-Create Solutions in Health Promotion with Users and Stakeholders? Front. Public Health 9:773907. doi: 10.3389/fpubh.2021.773907
18. Y.-Y. L. Andy, C.-P. Shen, Y.-S. Lin, H.-J. Chen, A.-C. Chen, L.-C. Cheng, T.-F. Tsai,C.-T. Huang, L.-M. Chuang, and F. Lai, “Continuous, personalized healthcare integrated platform,” in TENCON 2012 IEEE Region 10 Conference. IEEE, 2012, pp. 1–6.
19. M. Peleg, Y. Shahar, S. Quaglini, A. Fux, G. Garc´ıa-S´aez, A. Goldstein, M. E. Hernando,D. Klimov,I. Mart´ınez-Sarriegui, C. Napolitano et al., “Mobiguide: a personalized and patient-centric decision-support system and its evaluation in the atrial fibrillation and gestationaldiabetes domains,” User Modeling and User-Adapted Interaction, vol. 27, no. 2, pp.159–213, 2017.
20. 20.R. Hohemberger, C. E. da Roza, F. R. Pfeifer, R. M. da Rosa, P. S. S. de Souza, A. F. Lorenzon,M. C. Luizelli, and F. D. Rossi, “An approach to mitigate challenges to the electronic health records storage,”Measurement, p. 107424, 2020.
21. N. A. Busis, “How can i choose the best electronic health record system for my practice?”Neurology, vol. 75, no. 18 Supplement 1, pp. S60–S64, 2010.
22. A. L. Weathers and G. J. Esper, “How to select and implement an electronic health record in a neurology practice,” Neurology: Clinical Practice, vol. 3, no. 2, pp. 141–148, 2013.

23. E. M. Hart, P. Barmby, D. LeBauer, F. Michonneau, S. Mount, P. Mulrooney, T. Poisot, K. H.Woo, N. B. Zimmerman, and J. W. Hollister, "Ten simple rules for digital data storage,"PLoS computational biology, vol. 12, no. 10, 2016.
24. 24.G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal, "Good enough practices in scientific computing," PLoS computational biology, vol. 13, no. 6, 2017.
25. S. I. Khan and A. S. M. L. Hoque, "Towards development of health data warehouse: Bangladesh perspective," in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE, 2015, pp. 1–6.
26. T. K. Mackey, T.-T. Kuo, B. Gummadi, K. A. Clauson, G. Church, D. Grishin, K. Obbad,R. Barkovich, and M. Palombini, "'fit-for-purpose?'–challenges and opportunities for applications of blockchain technology in the future of healthcare," BMC medicine, vol. 17, no. 1, p. 68, 2019.
27. Saif Ur Rehman, Abdul Rehman Javed, Mohib Ullah Khan, Mubashar Nazar Awan, Adees Farukh & Aseel Hussien (2020) PersonalisedComfort: a personalised thermal comfort model to predict thermal sensation votes for smart building residents, Enterprise Information Systems, DOI: 10.1080/17517575.2020.1852316
28. Mubashar, A., Asghar, K., Javed, A. R., Rizwan, M., Srivastava, G., Gadekallu, T. R., & Shabbir, M. (2021). Storage and proximity management for centralized personal health records using an ipfs-based optimization algorithm. Journal of Circuits, Systems and Computers, 2250010.
29. Gadekallu, T.R., Khare, N., Bhattacharya, S. et al. Deep neural networks to predict diabetic retinopathy. J Ambient Intell Human Comput (2020). https://doi.org/10.1007/s12652-020-01963-7
30. Reddy, G.T., Reddy, M.P.K., Lakshmanna, K. et al. Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. Evol. Intel. 13, 185–196 (2020).
31. T. Trojer, B. Katt, T. Schabetsberger, R. Mair, and R. Breu, "The process of policy authoring of patient-controlled privacy preferences," in International Conference on Electronic Healthcare. Springer, 2011, pp. 97–104.
32. Analytics Vidhya.What is confusion matrix (2020).https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5 (Accessed November 17, 2020).

Figure 1.TIF



Patient Real Time IOT Health Data
Convert the data into Blocks
Expert Preference
Patient Preference
Preference Recommendation System
Store the Health Data in Suitable Health Repository

Figure 2.TIF

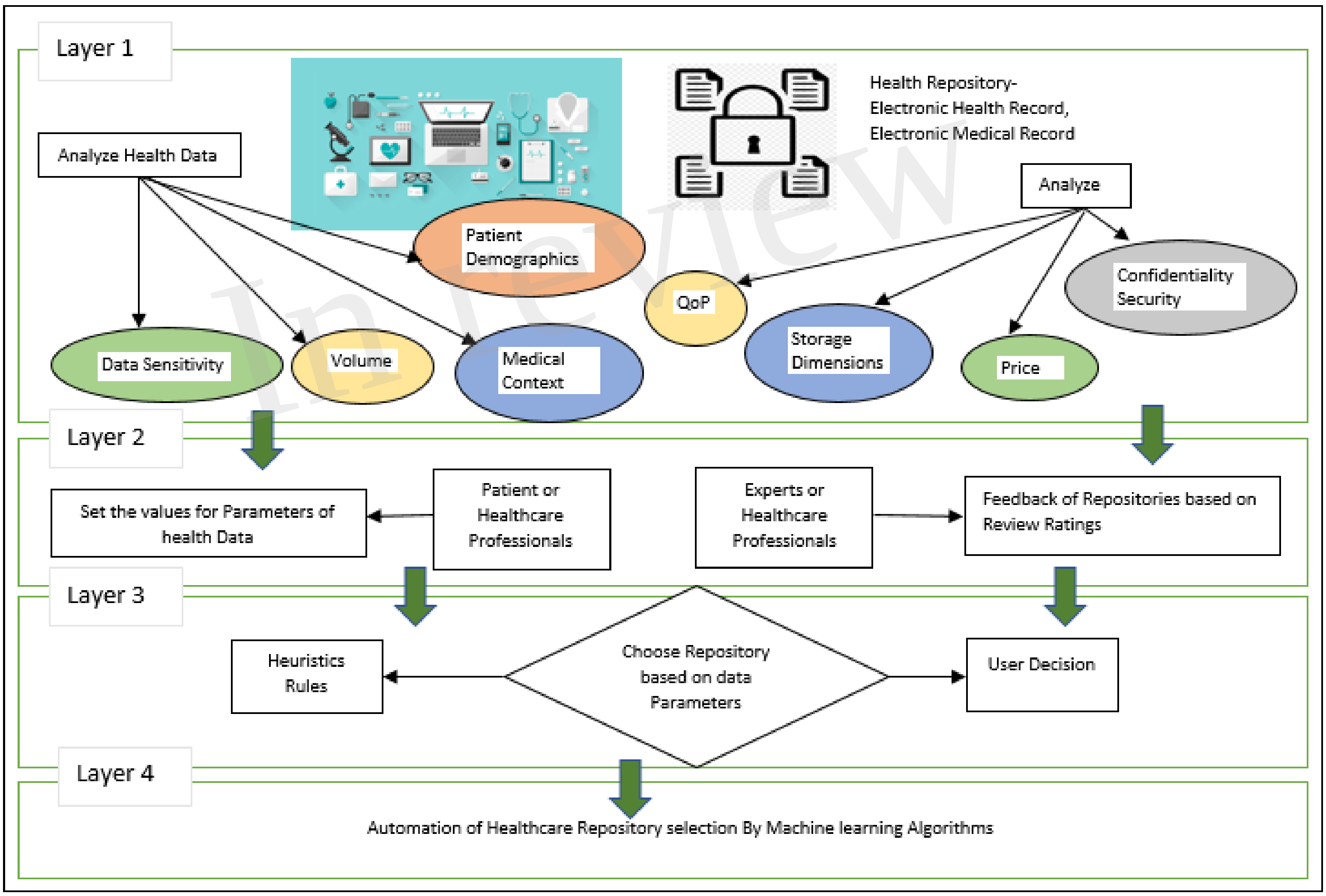

Layer 1
Analyze Health Data
Health Repository-
Electronic Health Record,
Electronic Medical Record
Analyze
Patient Demographics
Confidentiality Security
QoP
Data Sensitivity
Volume
Medical Context
Storage Dimensions
Price
Layer 2
Set the values for Parameters of health Data
Patient or Healthcare Professionals
Experts or Healthcare Professionals
Feedback of Repositories based on Review Ratings
Layer 3
Heuristics Rules
Choose Repository based on data Parameters
User Decision
Layer 4
Automation of Healthcare Repository selection By Machine learning Algorithms

Figure 3.TIF

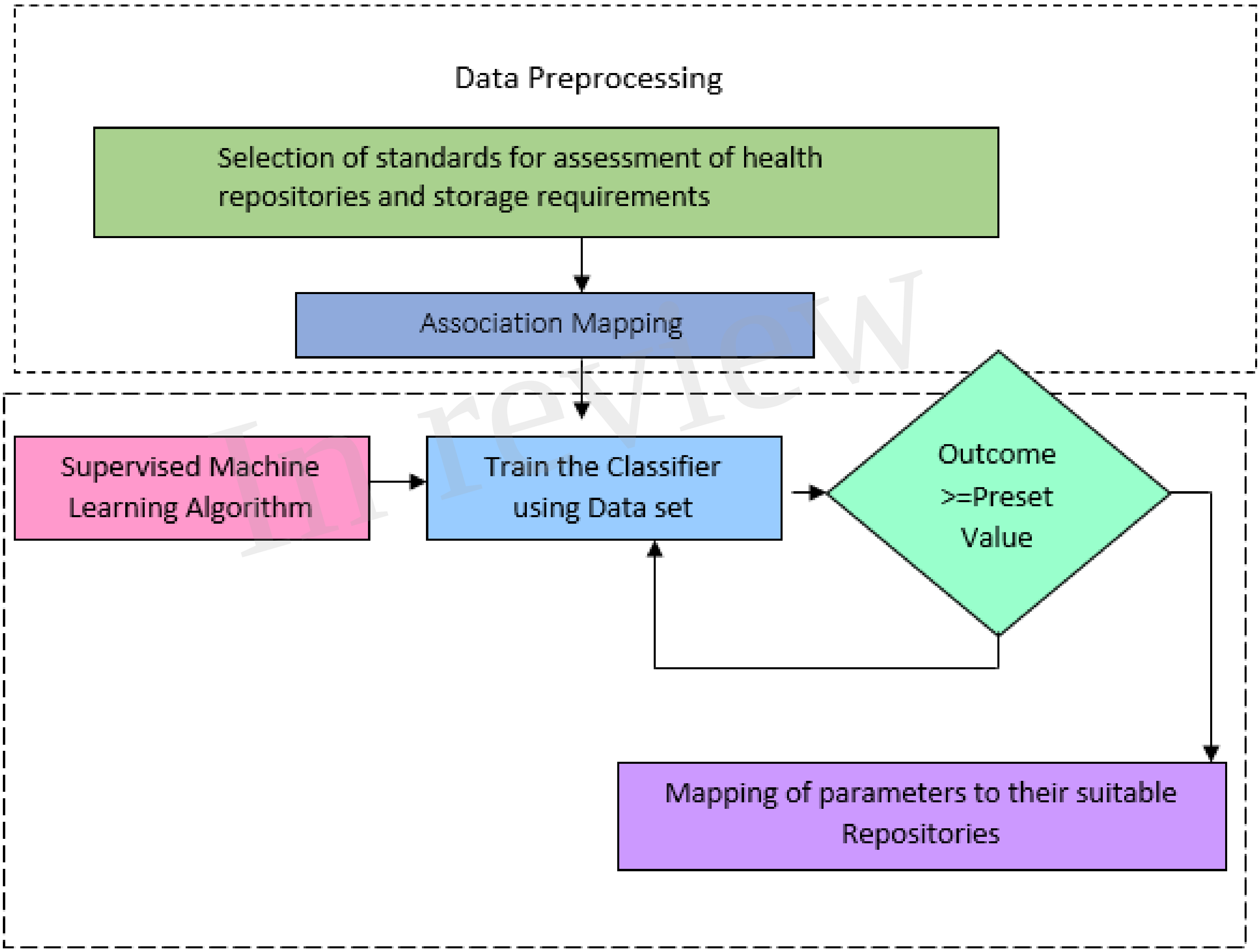
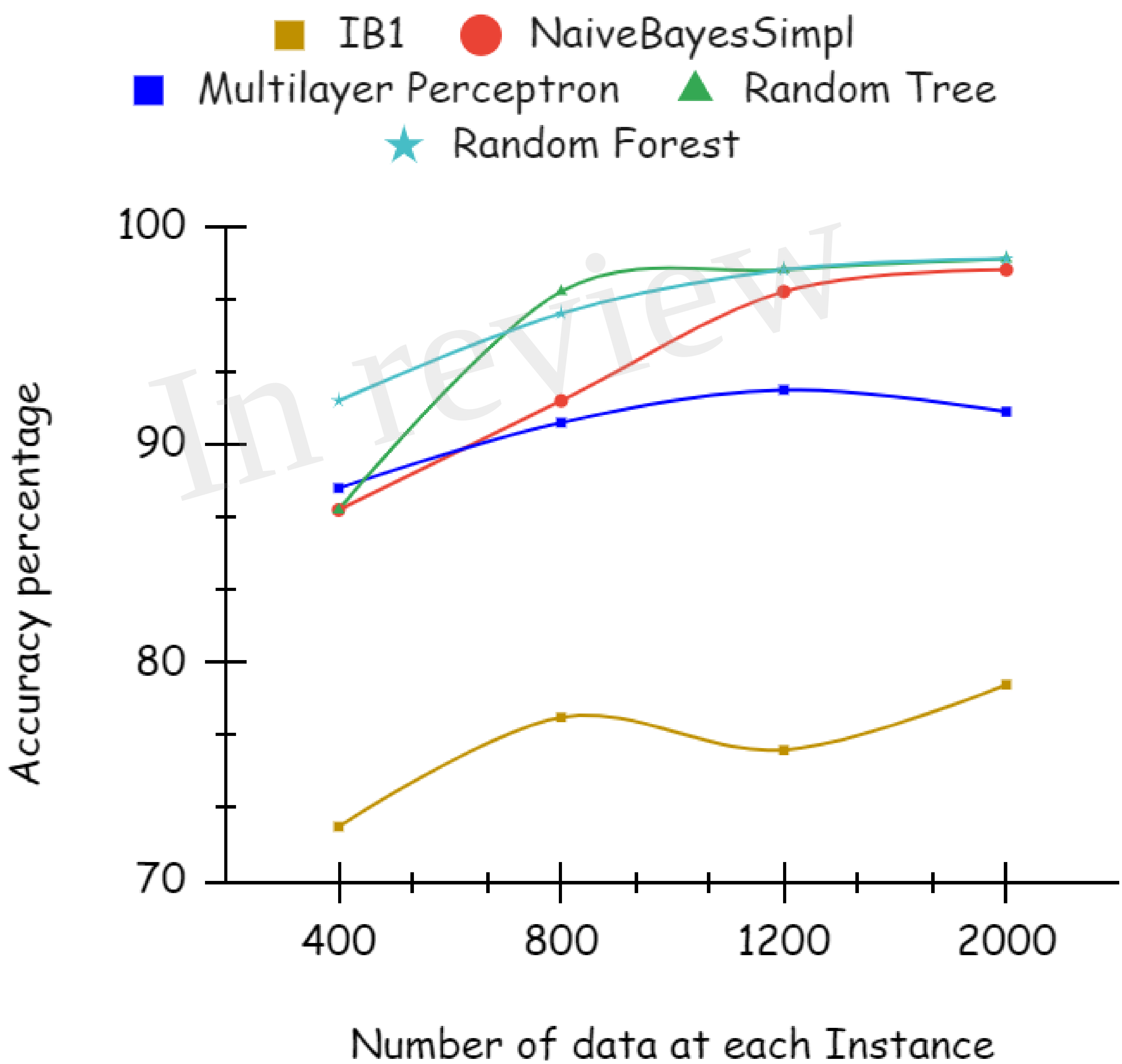

Data Preprocessing
Selection of standards for assessment of health repositories and storage requirements
Association Mapping
Supervised Machine Learning Algorithm
Train the Classifier using Data set
Outcome >=Preset Value
Mapping of parameters to their suitable Repositories

Figure 4.TIF



IB1
NaiveBayesSimpl
Multilayer Perceptron
Random Tree
Random Forest
100
90
80
70
Accuracy percentage
400
800
1200
2000
Number of data at each Instance
In review

Figure 5.TIF

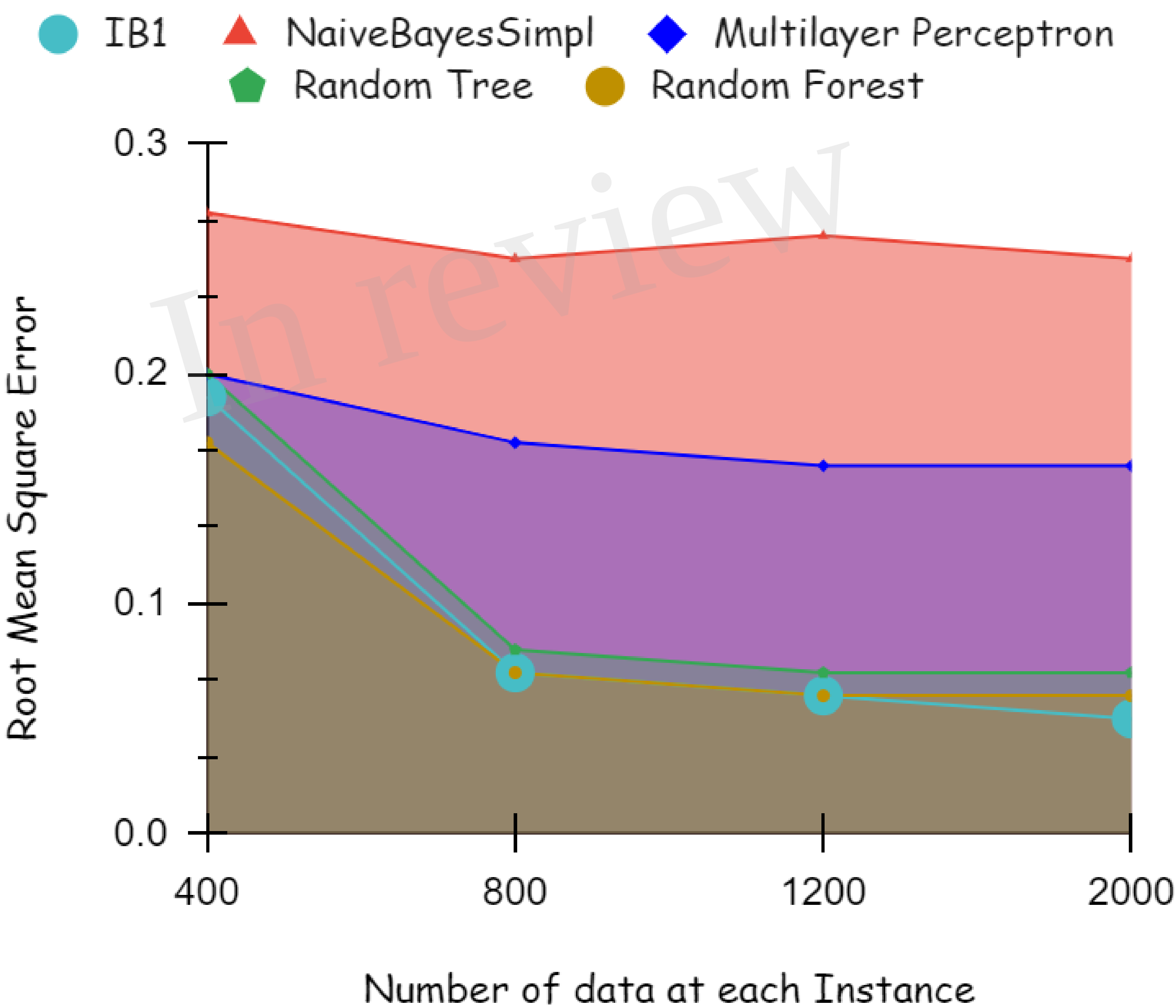
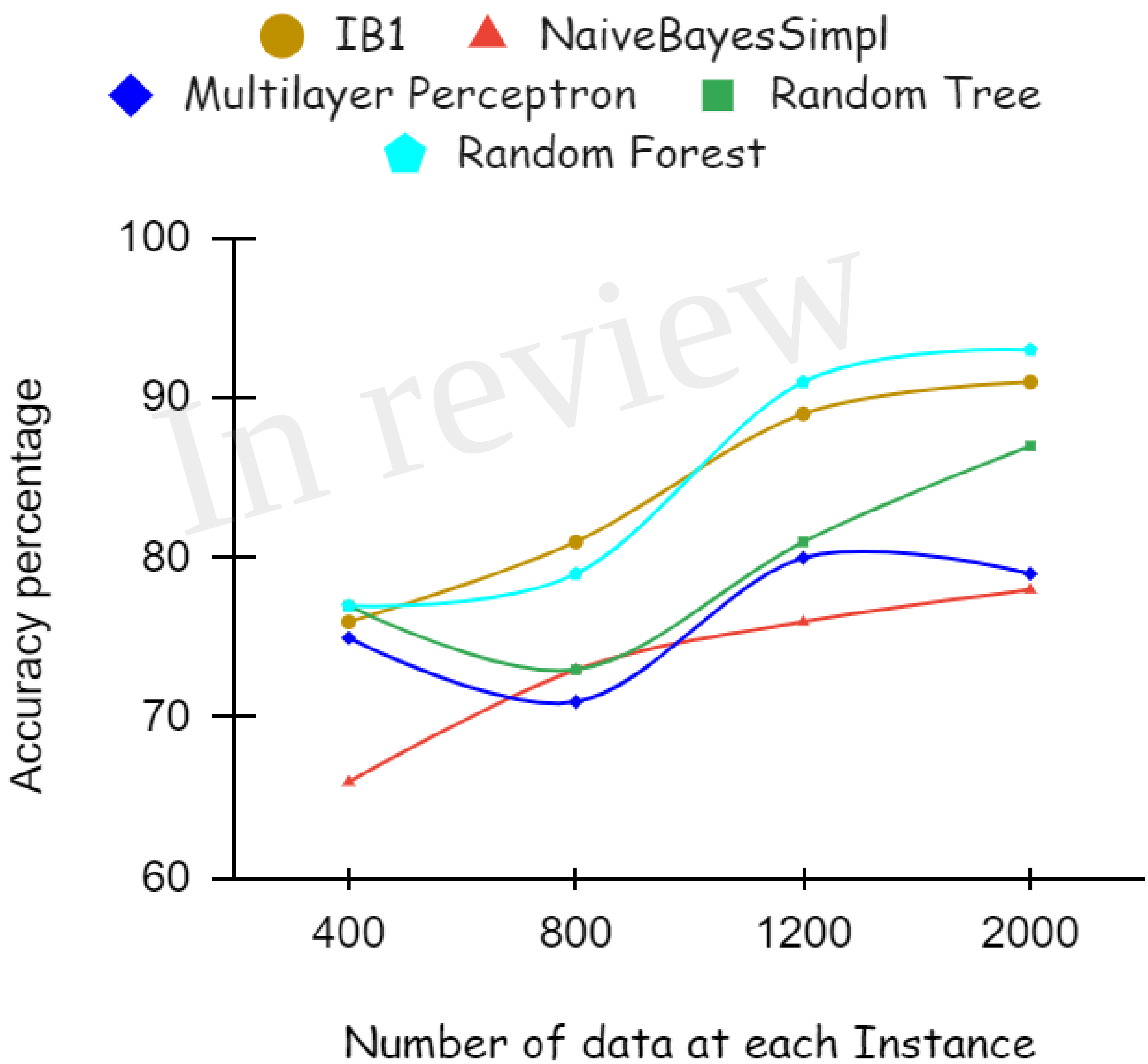

IB1
NaiveBayesSimpl
Multilayer Perceptron
Random Tree
Random Forest
0.3
0.2
0.1
0.0
Root Mean Square Error
400
800
1200
2000
Number of data at each Instance
In review

Figure 6.TIF



IB1
NaiveBayesSimpl
Multilayer Perceptron
Random Tree
Random Forest
100
90
80
70
60
Accuracy percentage
400
800
1200
2000
Number of data at each Instance
In review

Figure 7.TIF

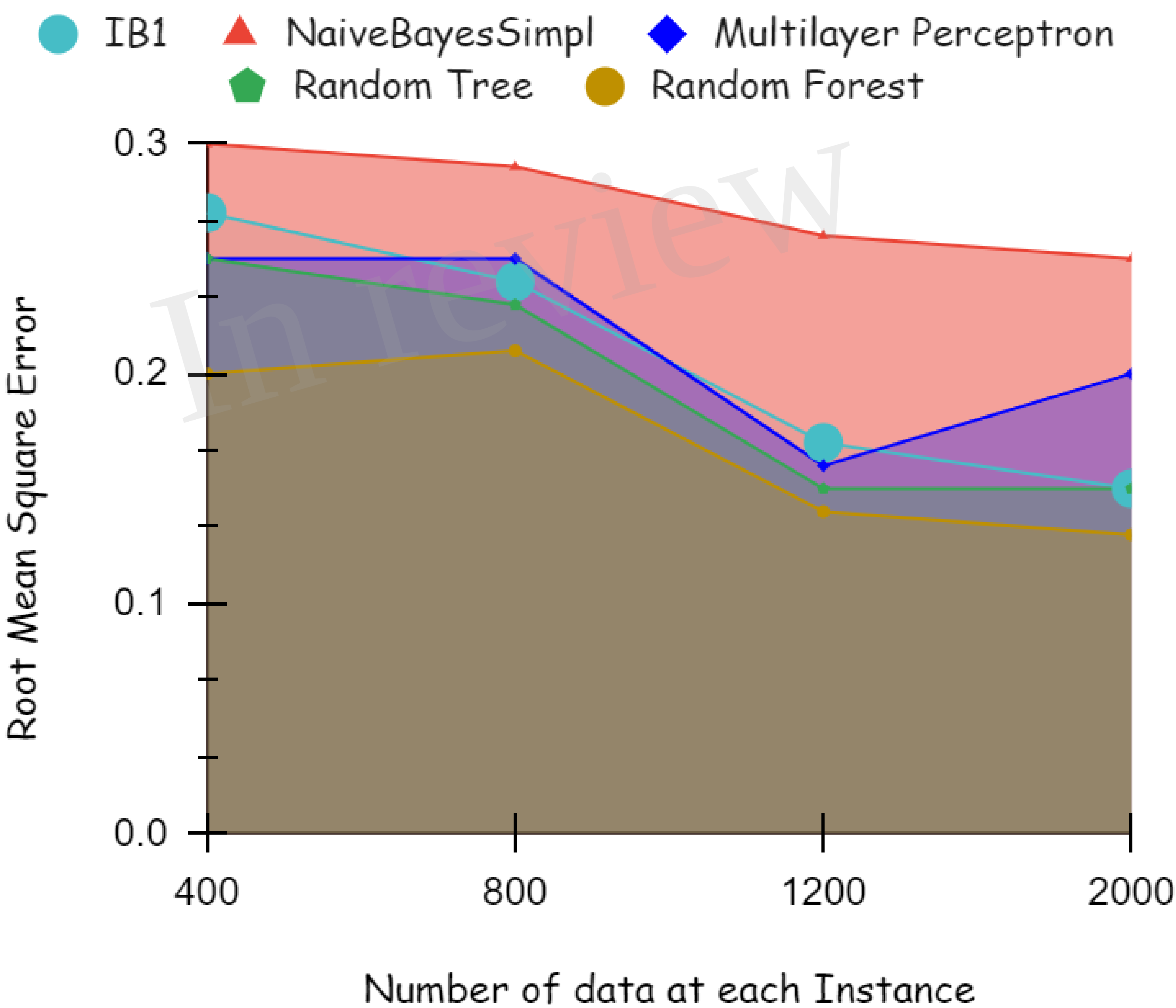

IB1
NaiveBayesSimpl
Multilayer Perceptron
Random Tree
Random Forest
0.3
0.2
0.1
0.0
Root Mean Square Error
400
800
1200
2000
Number of data at each Instance
In review

Figure 8.TIF

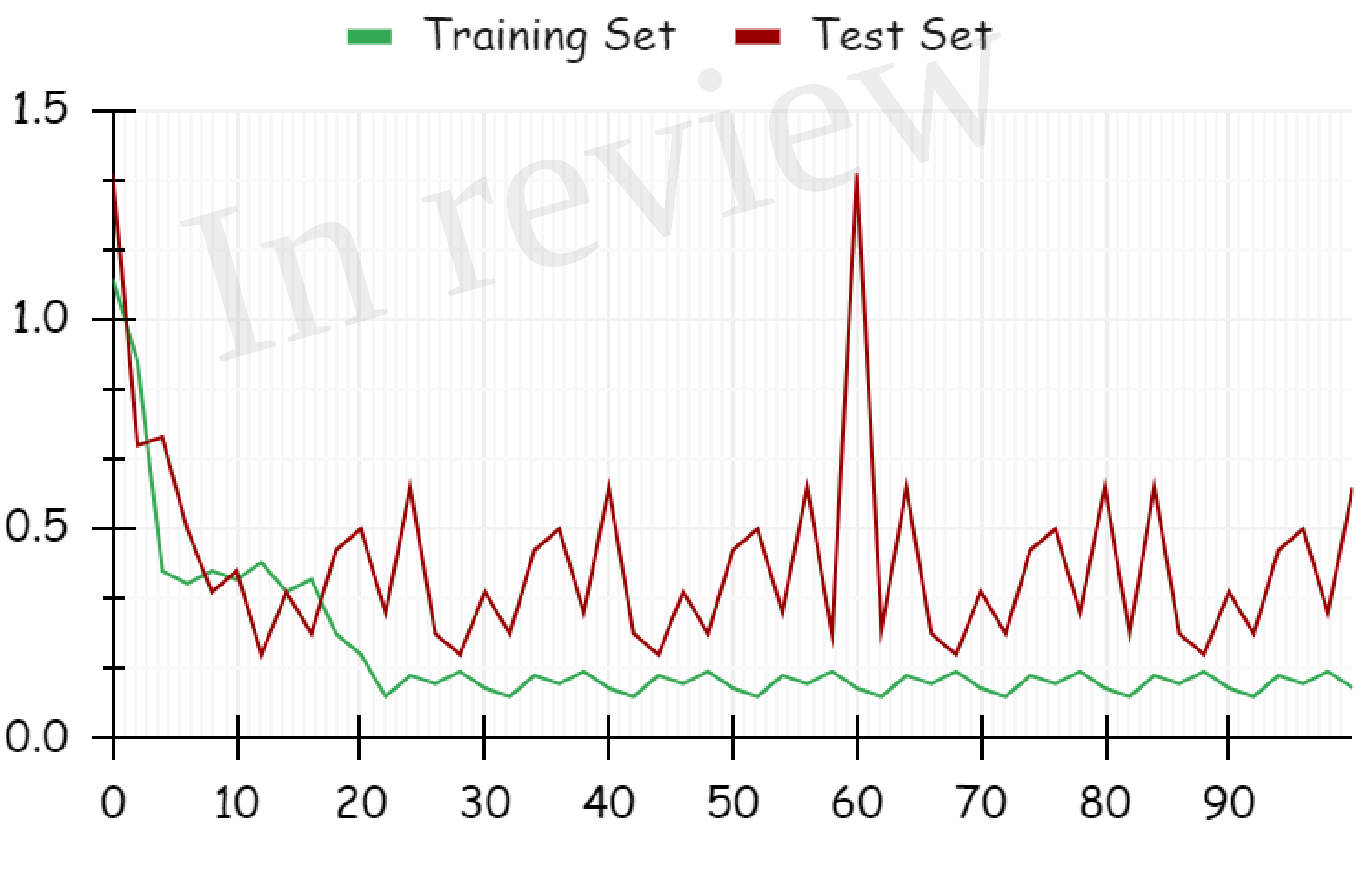


Figure 9.TIF

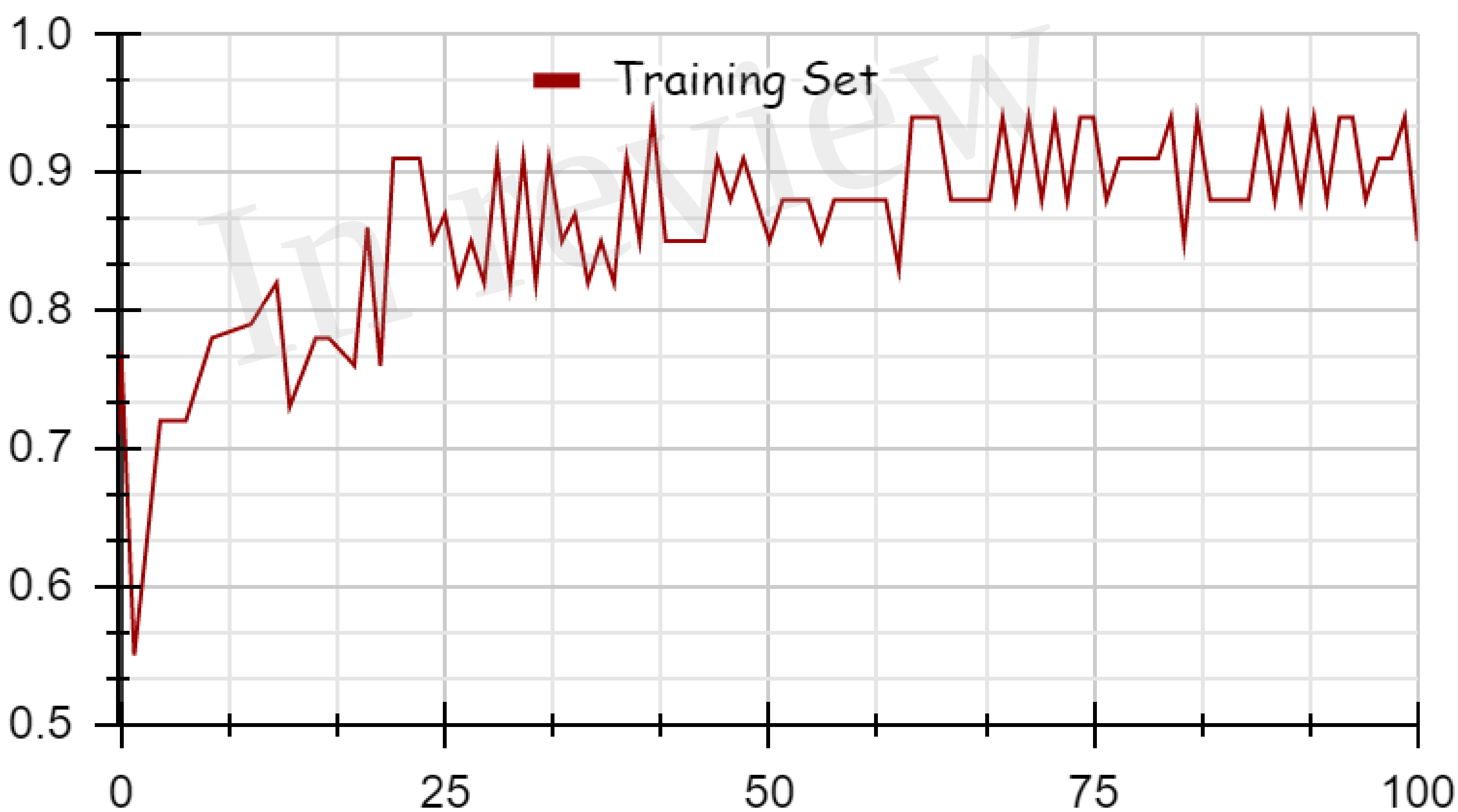


Figure 10.TIF

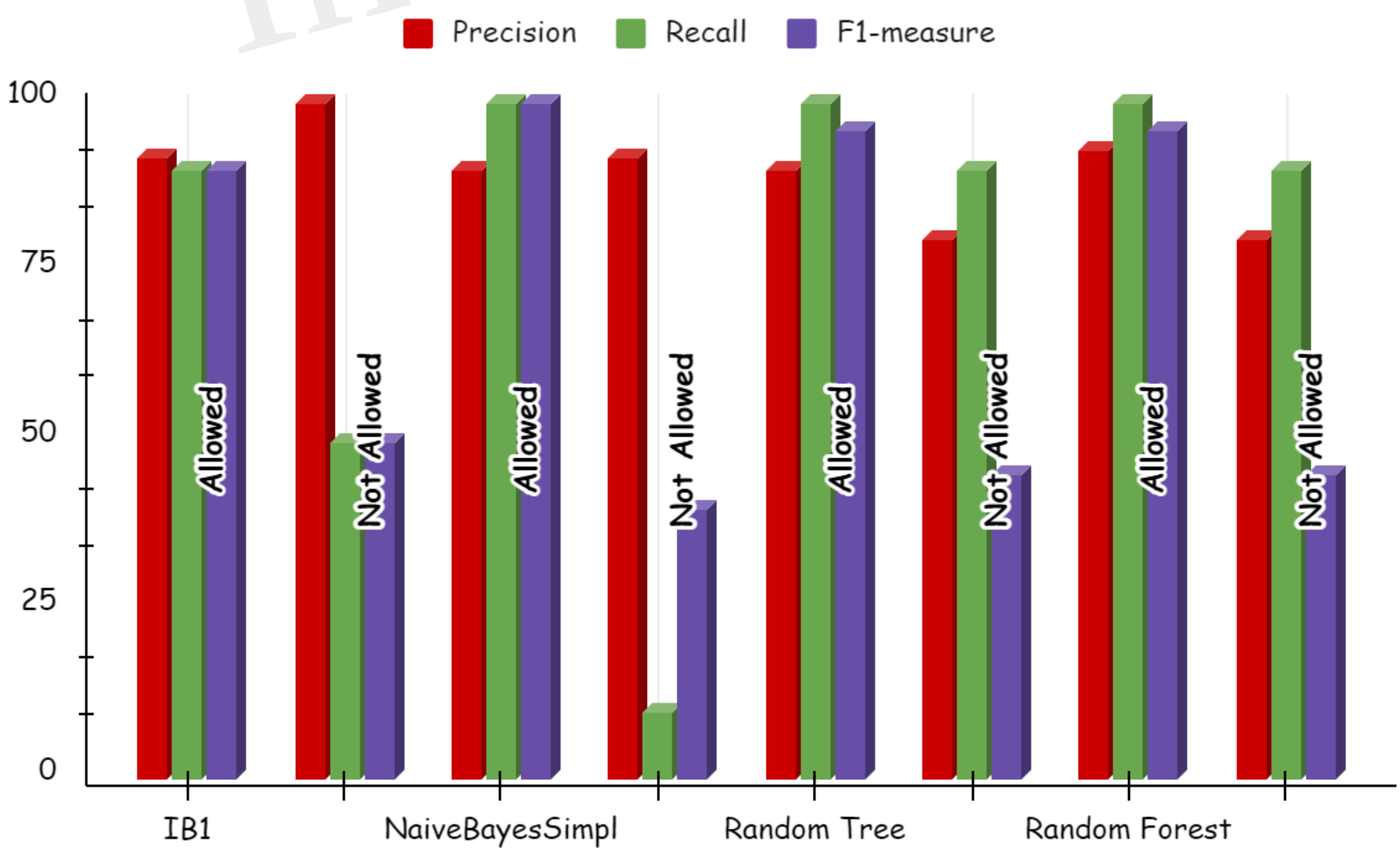


Figure 11.TIF

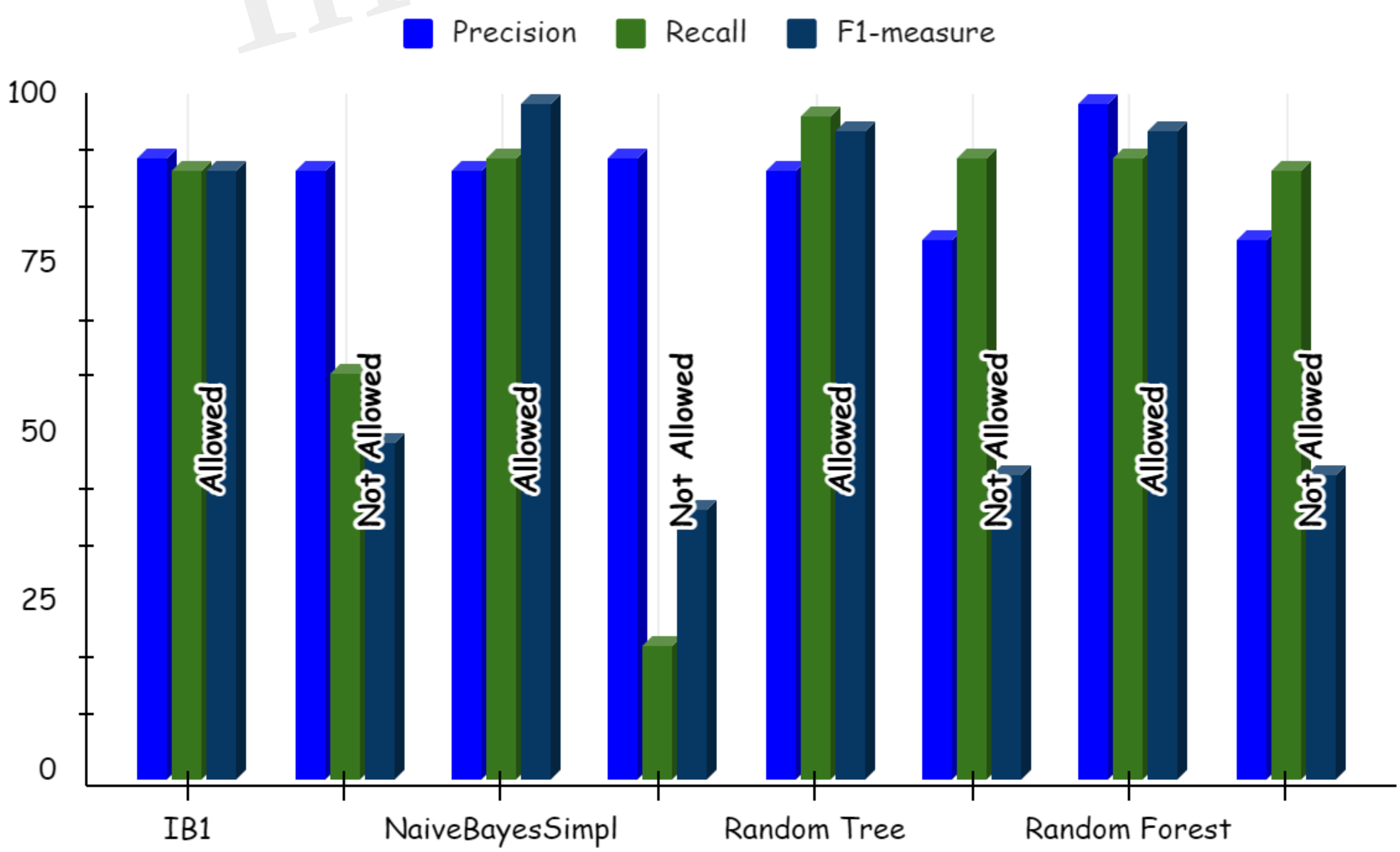


Figure 12.TIF