ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

PRACTICAL STRATEGIES FOR EXTREME MISSING DATA IMPUTATION IN DEMENTIA DIAGNOSIS

G. Uma Maheshwari, Professor, Department Of AIML, SICET, Hyderabad
Udigiri Rishika, Pyatlo Sai Kumar Reddy, Penugonda Krushitha, Goranta Shreya
UG Student Department Of AIML, SICET, Hyderabad

Abstract

Accurate computational models for clinical decision support systems require clean and reliable data but, in clinical practice, data are often incomplete. Hence, missing data could arise not only from training datasets but also test datasets which could consist of a single undiagnosed case, an individual. This work addresses the problem of extreme missingness in both training and test data by evaluating multiple imputation and classification workflows based on both diagnostic classification accuracy and computational cost. Extreme missingness is defined as having ~50% of the total data missing in more than half the data features. In particular, we focus on dementia diagnosis due to long time delays, high variability, high attrition rates and lack of practical data imputation strategies in its diagnostic pathway. We identified and replicated the extreme missingness structure of data from a real-world memory clinic on a larger open dataset, with the original complete data acting as ground truth. Overall, we found that computational cost, but not accuracy, varies widely for various imputation and classification approaches. Particularly, we found that iterative imputation on the training dataset combined with a reduced-feature classification model provides the best approach, in terms of speed and accuracy. Taken together, this work has elucidated important factors to be considered when developing a predictive model for a dementia diagnostic support system.*Index Terms*—Clinical decision support systems, medical expert systems, machine learning, missing data, data imputation, dementia, ADNI data, Alzheimer's disease classification, data quality

1. INTRODUCTION

The issue of missing data is one of the most ubiquitous concerns in data science [1]. This is particularly the case in clinical and medical data, which frequently has many missing values [2]–[4] (see Fig. 1a for a real-world, routine (i.e. not clinical trial) Alzheimer's disease (AD) dataset). In recent years, there has been increased effort to assure data quality and reusability, and to automate the processes of discovering and analysing data by publishing data annotations and analyticalthis work was supported by the European Union's INTERREG VA Programme, managed by the Special EU Programmes Body (SEUPB), and additional support by Alzheimer's Research UK (XD, MB, ST, PLM., KW-L), Ulster University Research Challenge Fund (XD, MB, ST, PLM, KW-L), and the Dr George Moore Endowment for Data Science at Ulster University (MB). The views and opinions expressed in this paper do not necessarily reflect those of the European Commission or the Special EU Programmes Body (SEUPB).NM, SL, XD, GP, MB and KW-L (k.wong-lin@ulster.ac.uk) are with the Intelligent Systems Research Centre, Ulster University. DPF is with Pharmacology and Therapeutics, School of Medicine, National University of Ireland Galway. ST is with Altnagelvin Area Hospital, Western Health and

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

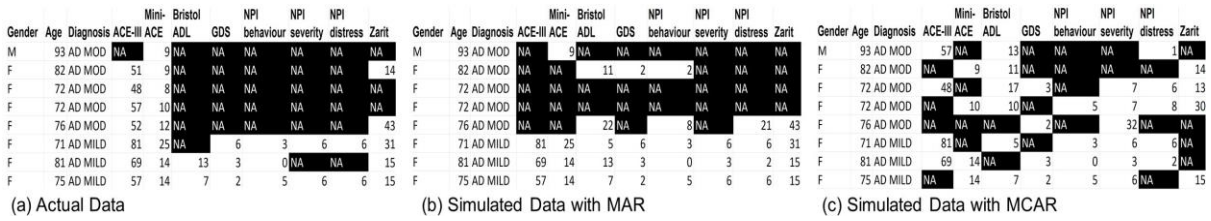
www.ijasem.org

Vol 18, Issue 1, 2024

Social Care Trust. PLM is with Ulster University, Northern Ireland Centre for Stratified Medicine, Biomedical Sciences Research Institute, Clinical Translational Research and Innovation Centre.See acknowledgements for ADNI below.

Data imputation strategies can further be divided into single imputation methods, in which a single estimate for the missing data is generated, and multiple imputation methods, which generate multiple estimates for each missing value and therefore will produce multiple imputed datasets for further analysis [2], [16]. Another crucial distinction is between supervised data imputation methods, where the class label is known, and unsupervised methods, which operate in the absence of a class label [17]. It is also useful to highlight that many commonly used imputation methods are iterative imputation methods which impute the entire dataset repeatedly until an optimum is reached e.g. [18], [19].

The appropriate strategy for dealing with missing data will depend to some extent on the type of missingness. Missing data is often categorized into three types: missing at random (MAR); missing completely at random (MCAR); and missing not at random (MNAR) [20]. In the case of MAR, the probability that data is missing depends upon the variables



Gender	Age	Diagnosis	ACE-III	Mini-ACE	Bristol ADL	GDS	NPI behaviour	NPI severity	NPI distress	Zarit
M	93	AD MOD	NA	9	NA	NA	NA	NA	NA	NA
F	82	AD MOD	51	9	NA	NA	NA	NA	NA	14
F	72	AD MOD	48	8	NA	NA	NA	NA	NA	NA
F	72	AD MOD	57	10	NA	NA	NA	NA	NA	NA
F	76	AD MOD	52	12	NA	NA	NA	NA	NA	43
F	71	AD MILD	81	25	NA	6	3	6	6	31
F	81	AD MILD	69	14	13	3	0	NA	NA	15
F	75	AD MILD	57	14	7	2	5	6	6	15

(a) Actual Data

Gender	Age	Diagnosis	ACE-III	Mini-ACE	Bristol ADL	GDS	NPI behaviour	NPI severity	NPI distress	Zarit
M	93	AD MOD	NA	9	NA	NA	NA	NA	NA	NA
F	82	AD MOD	NA	NA	11	2	2	NA	NA	NA
F	72	AD MOD	NA	NA	NA	NA	NA	NA	NA	NA
F	72	AD MOD	NA	NA	NA	NA	NA	NA	NA	NA
F	76	AD MOD	NA	NA	22	NA	8	NA	21	43
F	71	AD MILD	81	25	5	6	3	6	6	31
F	81	AD MILD	69	14	13	3	0	3	2	15
F	75	AD MILD	57	14	7	2	5	6	6	15

(b) Simulated Data with MAR

Gender	Age	Diagnosis	ACE-III	Mini-ACE	Bristol ADL	GDS	NPI behaviour	NPI severity	NPI distress	Zarit
M	93	AD MOD	57	NA	13	NA	NA	NA	1	NA
F	82	AD MOD	NA	9	11	NA	NA	NA	NA	14
F	72	AD MOD	48	NA	17	3	NA	7	6	13
F	72	AD MOD	NA	10	10	NA	5	7	8	30
F	76	AD MOD	NA	NA	NA	2	NA	32	NA	NA
F	71	AD MILD	81	NA	5	NA	3	6	6	NA
F	81	AD MILD	69	14	NA	3	0	3	2	NA
F	75	AD MILD	NA	14	7	2	5	6	NA	15

(c) Simulated Data with MCAR

Fig. 1. Sample Alzheimer's disease (AD) dataset from a memory clinic and its breakdown of data missingness. (a) Actual sample data

Data description

1) Medical data were edited to remove missing data featuresAnonymized medical data were e xtracted in CSV archive format from the Hospital Memory Assessment (WHSCT) of Altnage lvin Regional Hospital. Equity approval for this was obtained from the Office of Justice for N orthern Ireland (ORECNI, HSC REC B reference number: 17/NI/0142; IRAS scheme ID: 23 0077). These data were used to identify types of deletions in modern world medical records f or replication in the ADNI dataset. An example of the dataset is shown in Figure 1. 1 A. Ther e are 189 lines in total, and each line represents one patient. Cells with missing values appear in black. Features included 7 different Cognitive and Functional Assessment (CFA) scores, as well as gender, age, and text-based diagnostic information. AD Diagnosis

EcogPTTotal ECog (Patient) - Total Patients [57] 0.338375

Now missing from global actual treatment data as described in Section II.B.2. Rows with m issing values

for any of these properties were removed, creating an initial complete ADNIMERGE confi guration file (base file) containing 1185 lines, each row representing the participant. Becau se our original medical records are not long, multiple visits from the same participant at dif ferent times are considered variable. These data were class unbalanced and included 478 he

althy controls, 614 MCI, and 93 AD cases. This basic information provides the ground trut h for our research. From these data, significant missing data were removed and used for de cision making and classification assessment.

2) Missing data

Next, we looked for the relationship between missing valuesand the perceived decline of th e individual/patients. Although CFA in ADNIMERGE was not available in clinical data, pr evious studies have provided a correlation between the ACEIII score (in our clinical data) a nd the MMSE score (in ADNIMERGE) [60]. In particular, these two CFAs attempted to in clude missing patterns in clinical data in ADNIMERGE but were subsequently disregarded in the analysis (see below). We used ACEIII scores from the clinical literature as a basis fo r the association between withdrawal and cognitive impairment to support this study, witho ut using different outcome measures (which will be doubled in subsequent analysis). > Firs t, we regressed the proportion of missing values

in the clinical data set for ACEIII. The equation results (see Section III.1) are used to create missing data in the ADNIMERGE database. Specifically, MMSE scores in ADNIMERGE were converted to ACEIII scores using the conversion table in [60] . This transformation was used to add missing values to the CFA variable in the ADNIMERGE dataset. All rows are missing because this will not reflect the relationship between variables in thew data. In Section III.A, we show that the proportion of CFA values with missing data is very high. T herefore, a total of 10 synthetic ADNIMERGE datasets with different missing values were created to ensure good results. ACEIII and MMSE scores were excluded from subsequent a nalyzes because ACEIII was not included in ADNIMERGE and MMSE was not specificall y selected. Interventions [1] were used for analysis because they are easy to interpret and ca n be standardized. We also used a multivariate imputation method called predictive mean matching (PMM) [61] –

[64] from the Pursuit of Multivariate Chained Equations (MICE) package in R [65] . We u se PMM as one imputation (PMM1) and an average of 5, 10, 15, and 50 imputations (PMM 5, PMM10, PMM15, and PMM50, respectively). It should be noted that PMM is the defaul t method of MICE and often multiple installation packages are used. Imputation algorithms such as the KNN method [35] are commonly used on complete data and are not suitable fo r the majority of our missing data and are therefore ignored. i) Linearly regress the observe d values from each row onto the other rows to obtain the set of coefficients; (iii)< br>

Using newly created coefficients to create estimated values for missings; values

in this column (iv) Check with variables that they predict a value close to the predicted val ue for missing data; (v) A state is selected from these states and its observed consequences are assigned to replace missing values. Steps (ii) to (v) are repeated for each row and the en tire process is repeated 10 times to generate the estimated data. One assignment file was cr eated for PMM1, while 5 assignment files were created for PMM5 (see Supplementary Fig ure 1 for details).] 66] uses random forest (RF) regression to impute missing data [67]. The MissForest imputation method was chosen because it has been shown to outperform MICE [18], [68] in imputation, including some assumptions about missing data [18]. The MissFo rest method requires the following steps:

(i) Enter the column definition for each missing value in dataset D to create the imputed dat a D (ii) Copy D–

to D (iii) D for each row in D , use the row with no valueto create the RF model and use t

his model to predict missing values; , then output D-
if the maximum number of iterations has been reached, then output D-
; This issue has been well studied [27] and how the PCAbased approach affects the determi nation of the accuracy of variables has been investigated. Bayesian PCA is a method that u ses an iterative process similar to expectation optimization with a Bayesian model to estima te the eigenvalues of the underlying data (see Insert Figure 3 for details). Correct R2 using l inear regression of imputed values(full data) as a measure of imputation accuracy, with val ues
> 0 to 1 (worst to maximum, respectively). On average, minimum and maximum R2 values

were obtained for each of the 10 synthetic data sets. This method is also used to calculate t he average interpolation accuracy for each variable using the missForest algorithm.) norma lizes blood sugar

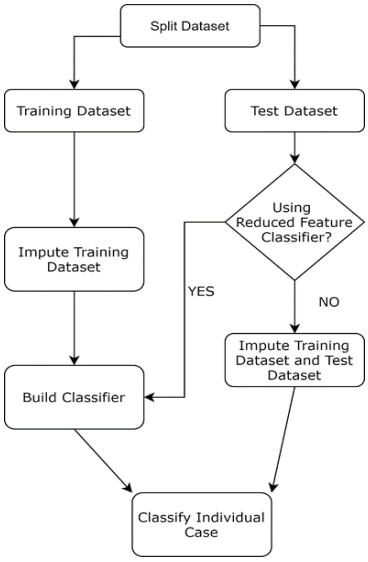
4) LOOCV validation
Use leave-one-out cross-
validation (LOOCV) [70] to evaluate the accuracy of the classification. According to LOO CV conditions, there is only one line in the test file. We use LOOCV to simulate patient dis tribution. LOOCV is also suitable for small files that may occur in some clinic/clinical envi ronments. Although LOOCV is considered expensive, it minimizes sample bias by using n early all data for each class while allowing prediction [71]. The methods we use to resolve missing values
in the test column can be divided into two groups: 1) Implement missing values
in the test column using the imputation method used for trainingInformation; or 2) using co unt reduction, where only nonmissing features of the test line are used to build the classific ation model. In a data set with N rows, the distribution model is created N times and tested sequentially on each row. The schematic diagram of this process is shown in Figure 2. The process shown in II is a variant of the general process shown in Figure 1. 2 (except workflo w H where interpolation is not used). The business process consists of a combination of dat a adjustment methods, data assignment methods, and classification methods. The RF classi fier (from the hat R package [72]) is used in most cases due to its generality and applicabili ty to many different datasets [18], while the SVM classifier (also from the hat package) is u sed in some cases. . Use a set of functions that test whether the assignment strategy has a di fferent effect on different objects. The Naive Bayes (NB) classifier (from the e1071 R pack age [73]) is used in (H) because it does not need a strategy to handle missing values; The pr ocessor can skip missing values while still using valuesfrom the same row of the data set. T he RF imputation method is used because it is the single most efficient imputation method, along with multiple imputation of PMM-

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

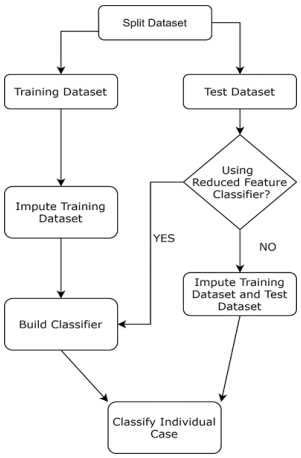
ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024



Final version of the assignment model with MissForest (see Supplementary Figure 2). For the PM5 and PMM-



Install Intel i7 processor, 16 GB RAM and R version 3.5.2. These tests are just a thread that a llows direct comparison of calculated values. The code is available at https://github.com/mac-n/BHImissingdata. ResultsSynthesizing missing data from clinical datTo better reduce the siz e of the ADNIMERGE dataset from realworld clinical data, we use data integration [55] for f eature selection. This algorithm selects the best features that influence the difference. betwee n the two (in our CDRSB score) and identify the 8 most important CFA characteristics. Table I shows the selected CFAs based on data sharing with categorical variables. Interestingly, mo st of the selected CFAs were completed by patient partners who attended study sites througho ut the ADNI study, rather than by the patients themselves (Table I, row 2). We then used the f irst 8 CFAs, as well as gender and age variables and categorical variables from the ADNIME RGE profile, to create our database, which is similar to the pain memory profile in the editor.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

We then checked for missing data in clinical memory to reconstruct the same missing data in the ADNIMERGE data. Culture of Addenbrooke's Cognitive Test (ACEIII). Although there were no different CFAs between clinical memory data and ADNIMERGE, MMSE scores we re included in ADNIMERGE, which may have caused the same type of deletions observed in our hospital memory data. Higher orders were tested, but higher orders were found to be non-significant in the polynomial decision (second order: pvalue = 0.051; third order: value = 0.3 9). 0.48 + (0.06 ACE-III), where Nmiss is the ratio of CFA values in each column and ACE-III is its normalized score. The constant 0.48 in the equation means that 48% of the CFA valu es are missing. The low p value (p = 2x116, n = 189) and low R2 (0.02502) of the regression indicate that cognitive decline (as measured by ACEIII scores) cannot be explained and there is a lot of missing information. This may be due to the failure of ACEIII as a discriminatory t ool for identifying severe and nonsevere pain in clinical data. Furthermore, based on our clini cal experience, only small gaps in the data are expected to be affected by cognitive impairme nt. Therefore the data can be considered MCAR or MAR. MMSE scores on the ADNI were c onverted to ACIII scores using a conversion table [60] . The above regression is used with the ACEIII score to generate the loss probability Pmiss, i. for each column in ADNIMERGE. Ea ch variable in each line I is replaced with the missing value byand the result is Pmiss, i. In thi s way, 10 missing data were generated from the entire ADNIERGE data, with the same level and type of missing data as our clinical data (see Section II.A.2). The imputation method is n ot more accurateAccording to the missing data, we made more imputation method. We found that the Proportional Proportion Method (PMM) and Random Forest (RF) methods gave the h ighest accuracy when tested against all data (ground truth) (Figure 3). The PMM assignment method is divided into PMM5, PMM10, PMM15, PMM50 (among 5, 10, 15 and 50 multiple rows, respectively). In particular, the PMM50 method is the most accurate method in recoveri ng the mean relative to the true value, with an average R2 of 0.86 across 10 synthetic data set s (Figure 3). This is not significantly (p = 0.204) higher than the accuracy when using the PM M15 imputation (mean 0.861), but is significantly higher than the accuracy of PMM10 (0.856) (ttest value for 10 data sets = 0.002). In contrast, the PMM15 method is significantly (p = 0. 001) greater than the RF method (mean 0.849), although RF is the only method with accuracy close to PMM. Therefore, the accuracy of PMM increases slightly when more decisions are made. All PMM methods provide 15 times greater accuracy than RF. The BC mean (average by rank) imputation method has reasonable accuracy (R2 = 0.735) for calculating simple met hods, but as an imputation method it has the disadvantage that it cannot be used to assign unk nown test lines. Finally, mean and mean methods do not provide accuracy. 3, gray bars), then we learned how to calculate the value of the individual assignment method. We see that diffe rent imputation methods have more computational time (Figure 3, black bars; note the logarit hmic scale). In particular, BPCA and PMM50 have similar timing, while RF is about twice as fast. It is twice as fast as PMM15 RF. Meaning, BC mean and standard deviation can be accu rate

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

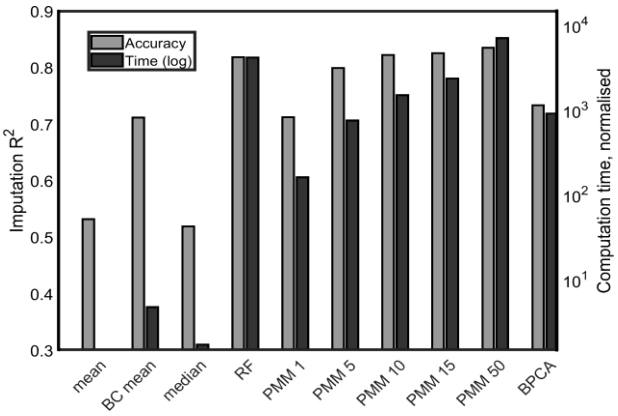


Fig. 3. Imputation accuracy R^2 and computation time depend on imputation methods. Missing data in test data will limit the use of many of the most popular methods, which are computationally expensive when the data is large. In this study, we replicated the i ncomplete model of the (memory) clinic of the modern world, focusing on the diagnosi s of AD.

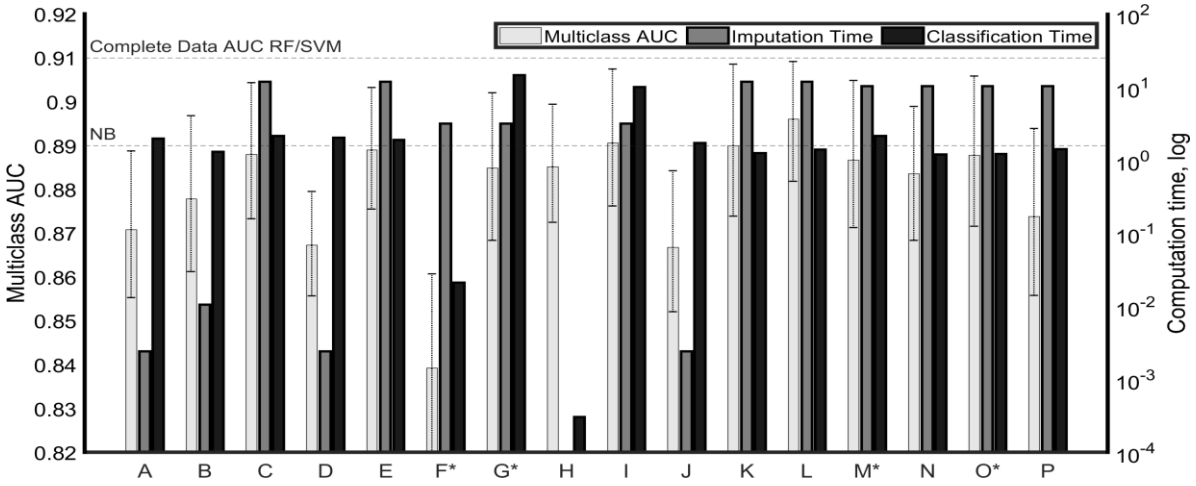


Fig. 4. Imputation and classification workflows evaluated for multiclass AUC (light grey, left axis; linear scale), imputation time and classification time (respectively dark grey and black, right axis; logarithmic scale.) Details of the workflows are explained in Table 2. Workflows marked with * impute the test dataset alongside the training dataset, with the test dataset class variable removed; hence imputation and classification must be performed together. Horizontal dashed lines: AUCs using complete dataset with RF and SVM (top), and with Naïve Bayes (NB) (bottom).

REFERENCES

[1] Sachin Kumar, Durga Toshniwal, "A data mining approach to characterize road accident locations", J. Mod. Transport. 24(1):62–72..

[2] Tessa K. Anderson, "Kernel density estimation and Kmeans clustering to profile road accident hotspots", Accident Analysis and Prevention 41,359–364.

[3] Shristi Sonal and Saumya Suman "A Framework for Analysis of Road Accidents" Proceedings of International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR)

[4] Analysis of road accidents in India using data mining classification algorithms- E. Suganya,S. Vijayrani.

[5] Hao, W., Kamga, C., Yang, X., Ma, J., Thorson, E., Zhong, M., & Wu, C., (2016), Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States, Transportation research part F: traffic psychology and behavior, 43, 379-386.

[6] Li, L., Shrestha, S., & Hu, G., (2017), Analysis of road traffic fatal accidents using data mining techniques, In Software Engineering Research, Management and Applications (SERA), IEEE 15th International Conference on (pp. 363-370). IEEE.

[7] El Tayeb, A. A., Pareek, V., & Araar, A. (2015). Applying association rules mining algorithms for traffic accidents in Dubai. International Journal of Soft Computing and Engineering.

[8] Bahram Sadeghi Bigham ,(2014),ROAD ACCIDENT DATA ANALYSIS: A DATA MINING APPROACH, Indian Journal Of Scientific Research 3(3):437-443.

[9] Divya Bansal, Lekha Bhambhu, "Execution of Apriori algorithm of data mining directed towards tumultuouscrimes concerningwomen", International Journal of AdvancedResearch in Computer Science and Software Engineering, vol. 3, no. 9, September 2013.

[10]S. Krishnaveni, M. Hemalatha, "A perspective analysis of traffic accident using data mining techniques", International Journal of Computer Applications, vol. 23, no. 7, pp. 40-48, June 2011.