Binary Journal of Data Mining & Networking 4 (2014) 54-57
http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

Data Mining Techniques in Cancer Research Area

Shiv Shakti Shrivastava*[1], Dr. Anjali Sant[2]

*Corresponding author:

Shiv Shakti Shrivastava

[1]Research Scholar, Mewar University, Chittorgarh (Raj.)
[2]Professor, BITS, Bhopal

Abstract

In this paper we present an analysis of the prediction of survivability on different attributes, rate of breast cancer patients using data mining techniques. The data used is the real data. The preprocessed data set, which have all the available twelve fields from the database. We have investigated data mining techniques:

Keywords: Breast cancer survivability, data mining.

Introduction

Cancer is a class of diseases characterized by out-of-control cell growth. There are over 100 different types of cancer, and each is classified by the type of cell that is initially affected.

Cancer harms the body when damaged cells divide uncontrollably to form lumps or masses of tissue called tumors (except in the case of leukemia where cancer prohibits normal blood function by abnormal cell division in the blood stream). Tumors can grow and interfere with the digestive, nervous, and circulatory systems and they can release hormones that alter body function. Tumors that stay in one spot and demonstrate limited growth are generally considered to be benign.

Advances in cancer medicine have traditionally come from detailed understanding of biological processes, later translated into therapeutic interventions, whose effectiveness is established by rigorous analysis of clinical trials. Over the last two decades the increasing throughput of data from microarray screening, spec- tral imaging and longitudinal studies are turning the under- standing of cancer pathology into as much a data-based as a biologically and clinically driven science, with potential to impact more strongly on evidence-based decision support moving towards personal-ized medicine [1]. This article is not intended as a comprehensive survey of data mining sequencing for monitoring genetic changes in tumor cells as they progress from normal to invasive [2].

Today, in the India, approximately one in eight women over their lifetime has a risk of developing breast cancer. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis [1].

The cause of cancer is due to irregular life style of human being. We found that the discovery of the survival rate or survivability of a certain disease is possible by extracting the knowledge from the data related to that disease. One of these data sources Surveillance Epidemiology and End Results), which is a unique, reliable and essential resource for investigating the different aspects of cancer. The SEER database combines patient-level information on cancer site, tumor pathology, stage, and cause of death [3, 4].

Related Work

A literature survey showed that there have been several studies on the survivability prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and survivability using data mining approaches like decision trees [7, 8, 9]. Accuracy. After a careful analysis of the breast cancer data used in [9], we have noticed that the number of “not survived” patients used does not match the number of “not alive” (field VSR) patients in the first 60 months of survival time. As a matter of fact, the number of “not survived” patients is expected to be around 20% based on the breast cancer survival statistics of 80% [1].

Methodology

In this paper, we have investigated classification data mining techniques. We are using real data from the Cance Research Hospital. In this paper, we used data and after selection of the criteria from different attributes to predict the survivability rate of breast cancer data set. We selected these three classification techniques to find the most suitable one for predicting cancer survivability rate.

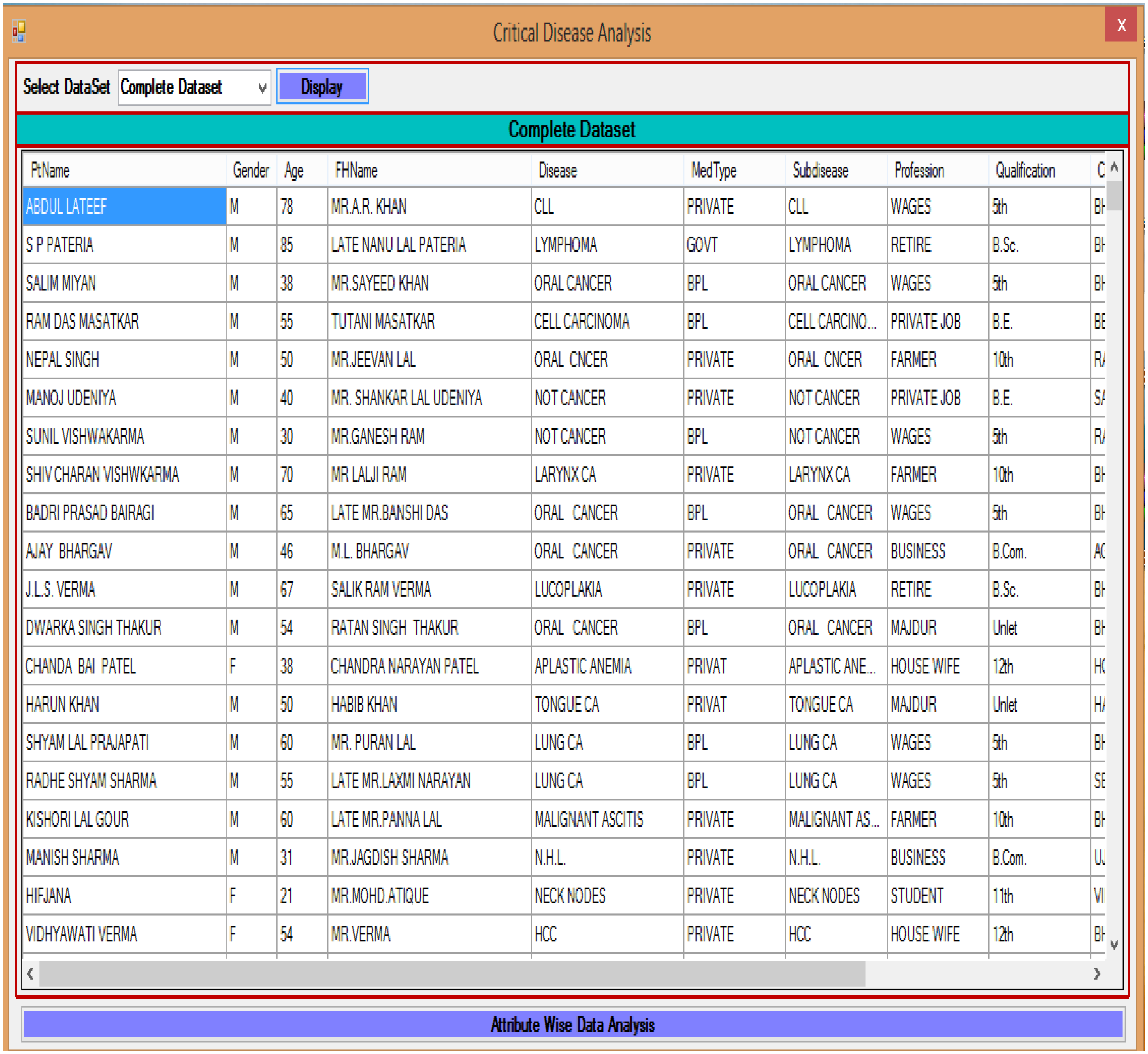
Feature Extraction

Images usually have a huge number of features. It is important to recognize and extract interesting features for an exacting task in order to decrease the complexity of processing. Not all the attributes of an image are useful for knowledge extraction. Image processing algorithm used, which automatically extract image attributes such as local color, global color, texture, structure. The extraction of the features from an image can finished using a variety of image processing techniques. Based on this, the image is processed to look for a measurement that helps in selecting the pixels that correspond to the centers of the nodule. We localize the extraction process to very small regions in order to ensure that we capture all areas [3].

Experimental Results

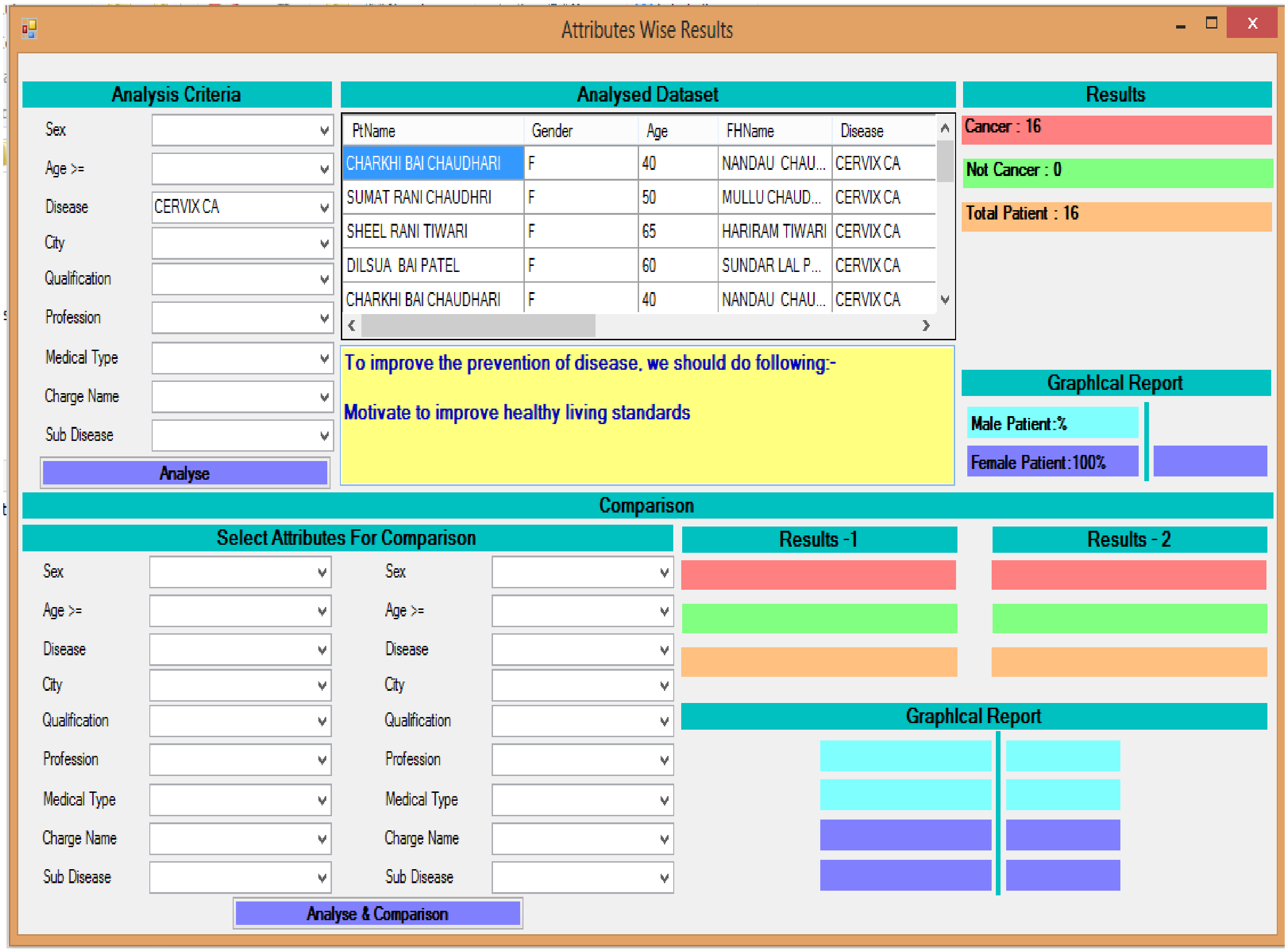
This study data mining techniques is compared and goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field these obtained results in this work differ from the

We are producing some of the snap shot of that made software and analytical analysis by weak software.



PtName	Gender	Age	FHName	Disease	MedType	Subdisease	Profession	Qualification	C
ABDUL LATEEF	M	78	MR.A.R. KHAN	CLL	PRIVATE	CLL	WAGES	5th	BH
S P PATERIA	M	85	LATE NANU LAL PATERIA	LYMPHOMA	GOVT	LYMPHOMA	RETIRE	B.Sc.	BH
SALIM MIYAN	M	38	MR.SAYEED KHAN	ORAL CANCER	BPL	ORAL CANCER	WAGES	5th	BH
RAM DAS MASATKAR	M	55	TUTANI MASATKAR	CELL CARCINOMA	BPL	CELL CARCINO...	PRIVATE JOB	B.E.	BE
NEPAL SINGH	M	50	MR.JEEVAN LAL	ORAL CNCER	PRIVATE	ORAL CNCER	FARMER	10th	RA
MANOJ UDENIYA	M	40	MR. SHANKAR LAL UDENIYA	NOT CANCER	PRIVATE	NOT CANCER	PRIVATE JOB	B.E.	SA
SUNIL VISHWAKARMA	M	30	MR.GANESH RAM	NOT CANCER	BPL	NOT CANCER	WAGES	5th	RA
SHIV CHARAN VISHWKARMA	M	70	MR LALJI RAM	LARYNX CA	PRIVATE	LARYNX CA	FARMER	10th	BH
BADRI PRASAD BAIRAGI	M	65	LATE MR.BANSHI DAS	ORAL CANCER	BPL	ORAL CANCER	WAGES	5th	BH
AJAY BHARGAV	M	46	M.L. BHARGAV	ORAL CANCER	PRIVATE	ORAL CANCER	BUSINESS	B.Com.	AC
J.L.S. VERMA	M	67	SALIK RAM VERMA	LUCOPLAKIA	PRIVATE	LUCOPLAKIA	RETIRE	B.Sc.	BH
DWARKA SINGH THAKUR	M	54	RATAN SINGH THAKUR	ORAL CANCER	BPL	ORAL CANCER	MAJDUR	Unlet	BH
CHANDA BAI PATEL	F	38	CHANDRA NARAYAN PATEL	APLASTIC ANEMIA	PRIVAT	APLASTIC ANE...	HOUSE WIFE	12th	HC
HARUN KHAN	M	50	HABIB KHAN	TONGUE CA	PRIVAT	TONGUE CA	MAJDUR	Unlet	HA
SHYAM LAL PRAJAPATI	M	60	MR. PURAN LAL	LUNG CA	BPL	LUNG CA	WAGES	5th	BH
RADHE SHYAM SHARMA	M	55	LATE MR.LAXMI NARAYAN	LUNG CA	BPL	LUNG CA	WAGES	5th	SE
KISHORI LAL GOUR	M	60	LATE MR.PANNA LAL	MALIGNANT ASCITIS	PRIVATE	MALIGNANT AS...	FARMER	10th	BH
MANISH SHARMA	M	31	MR.JAGDISH SHARMA	N.H.L.	PRIVATE	N.H.L.	BUSINESS	B.Com.	UJ
HIFJANA	F	21	MR.MOHD.ATIQUE	NECK NODES	PRIVATE	NECK NODES	STUDENT	11th	VI
VIDHYAWATI VERMA	F	54	MR.VERMA	HCC	PRIVATE	HCC	HOUSE WIFE	12th	BH

Figure 1. Attributes based table.



Disease: CERVIX CA

PtName	Gender	Age	FHName	Disease
CHARKHI BAI CHAUDHARI	F	40	NANDAU CHAU...	CERVIX CA
SUMAT RANI CHAUDHRI	F	50	MULLU CHAUD...	CERVIX CA
SHEEL RANI TIWARI	F	65	HARIRAM TIWARI	CERVIX CA
DILSUA BAI PATEL	F	60	SUNDAR LAL P...	CERVIX CA
CHARKHI BAI CHAUDHARI	F	40	NANDAU CHAU...	CERVIX CA

Cancer : 16

Not Cancer : 0

Total Patient : 16

To improve the prevention of disease, we should do following:-

Motivate to improve healthy living standards

Male Patient:%

Female Patient:100%

Figure 2. Attributes wise result

Conclusions and Future Work

This paper has outlined, discussed and techniques for the problem of breast cancer survivability prediction in database. Unlike the preclassification process used in [9]. The experimental results show that our approach outperforms the approach used in [9].

This study clearly shows that the preliminary results are promising for the application of the data mining methods into the survivability prediction problem in medical databases. Our analysis does not include records with missing data; future work will include the missing data in the EOD field from the old EOD fields prior to 1988. This might increase the performance as the size of the data set will increase considerably.

Finally, we would like to try survival time prediction of certain cancer data where the survivability is seriously low. We think of discrediting the survival time in terms of one year and then classifying using the aforementioned data mining algorithms.

References

[1]. International Journal of Advanced Research in Computer Science and Software Engineering "A Study of Detection of Lung Cancer Using Data Mining Classification Techniques" Volume 3, Issue 3, March 2013, ISSN: 2277 128X

[2]. American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

[3]. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Public-Use

Data (1973-2002), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2005, based on the November 2004 submission.

[4]. Cox DR. Analysis of survival data. London: Chapman & Hall; 1984.

[5]. Benjamin F. Hankey, et. al. The Surveillance, Epidemiology, and End Results Program: A National Resource. Cancer Epidemiology Biomarkers & Prevention 1999; 8:1117-1121.

[6]. Houston, Andrea L. and Chen, et. al.. Medical Data Mining on the Internet: Research on a Cancer Information System. Artificial Intelligence Review 1999; 13:437-466.

[7]. Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine 2002; 26:1-24.

[8]. Zhou ZH, Jiang Y. Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble. IEEE Trans Inf Technol Biomed. 2003 Mar; 7(1):37-42.

[9]. Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology 1999; 57:281-6.

[10]. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.

[11]. Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005.

[12]. Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, CA:Morgan Kaufmann; 1993.

[13]. Madeira SC, and Oliveira AL. "Biclustering al-gorithms for biological data analysis: A survey," *IEEE/ ACM Trans. Computat. Biol. Bioinform.*, vol. 1, no. 1, pp. 24–45, 2004.

[14]. Carrivick L. et al., Identification of prognostic sig-natures in breast cancer microarray data using Bayesian techniques," *J. R. Soc. Interface*, vol. 3, no. 8, pp. 367–381, 2006.

[15]. Ben-Hur, A. Elisseeff, and I. Guyon, "A stabil-ity based method for discovering structure in clus-tered data," in *Biocomputing (Proc. Pacific Symp.)*, vol. 7, R. B. Altman and K. Lauderdalc, Eds. Kauai, Hawaii, USA, 2002, pp. 6–17.

[16]. Kerr MK, and Churchill GA. "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments," *Proc. Natl. Acad. Sci.*, vol. 98, pp. 8961–8965, 2001.

[17]. Kuncheva LI, and Vetrov DP. "Evaluation of sta-bility of k-means cluster ensembles with respect to ran-dom initialization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 11, pp. 1798–1808, 2006.

[18]. Gionis, H. Mannila, and P. Tsaparas, "Cluster-ing aggregation," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, 2007.

[19]. Nguyen N, and Caruana R. "Consensus clustering," in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, 2007, pp. 607–612.

[20]. Caruana R, Elhawary M, Nguyen N, and Smith C. "Meta clustering," in *Proc. 6th IEEE Int. Conf. Data Mining (ICDM)*, 2006, pp. 107–118.

[21]. Ciaramella, et al., "Interactive data analysis and clustering of genomic data," *Neural Netw.*, vol. 21, pp. 368–378, 2008.

[22]. Lee J A, and Verleysen M. *Nonlinear Dimensionality Reduction*. New York: Springer-Verlag, 2007.

[23]. Vesanto J. "SOM-based data visualization methods," *Intell. Data Anal.*, vol. 3, pp. 111–126, 1999.

[24]. Lisboa PJG, Ellis IO, Green AR, Ambrogi F, and Dias MB. "Cluster-based visualisation with scat-ter matrices," *Pattern Recogn. Lett.*, vol. 29, no. 13, pp. 1814–1823, 2008.

[25]. Hasty, D. McMillen, F. Isaacs, and J. J. Collins, "Com-putational studies of gene regulatory networks: In numero molecular biology," *Nature Rev. Gene.*, vol. 2, pp. 268–279, 2001.

[26]. de Jong H. "Modeling and simulation of genetic regulatory systems: A literature review," *J. Computat. Biol.*, vol. 9, no. 1, pp. 67–103, 2002.

[27]. Borowski EJ, and Borwein JM. *Computational Modeling of Genetic and Biochemical Networks*. Cambridge, MA: MIT Press, 2001.

[28]. Friedman N. "Inferring cellular networks using prob-abilistic graphical models," *Science*, vol. 303, p. 799, 2004.

[29]. Needham, J. Bradford, A. Bulpitt, and D. West-head, "A primer on learning in Bayesian networks for computational biology," *PLOS Computat. Biol.*, vol. 3, no. 8, pp. 1409–1416, 2007.

[30]. Gat-Viks, A. Tanay, D. Raijaman, and R. Shamir, "A probabilistic methodology for integrating knowledge and experiments on biological networks," *J. Computat. Biol.*, vol. 13, no. 2, pp. 165–181, 2006.

[31]. Schafer and K. Strimmer, "An empirical Bayes ap-proach to inferring large-scale gene association networks," *Bioinformatics*, vol. 21, no. 6, pp. 754–764, 2005.

[32]. Basso, et al., "Reverse engineering of regulatory networks in human B cells," *Nature Gene.*, vol. 37, no. 4, pp. 382–290, 2005.