Journal is aimed to publish research articles on theoretical foundations of information and computation, and of practical techniques for their implementation and application in computer systems. Computer Reviews is scientific, international peer-reviewed/Refereed online journal that is devoted to fields of computing to provide rapid publication of articles frequently in its issues. Computer Reviews is the place for exchange of information and research results within all areas of Computing and allied fields. Journal is aimed to publish research articles on theoretical foundations of information and computation, and of practical techniques for their implementation and application in computer systems.

**Peer Review process** : Journal uses double blind review system to ensure originality, timeliness, relevance, and readability.
Chief Editor

Dr. Hari Mohan Srivastava  
Ph.D., D.Sc. (h.c), D.Sc. (h.c)  
Professor Emeritus, Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia V8W 3R4, Canada  
http://www.math.uvic.ca/faculty/harimsri/  
harimsri@math.uvic.ca

Editors

Prof. Douglas Lyon, PhD, PE  
Fairfield University, Milford, CT 06461, USA  
Lyon@docjava.com  

Dr. Balakrushna Tripathy  
Senior Professor School of Computing Sciences and Engineering VIT, Vellore-632 014, Tamil Nadu, INDIA  
http://www.vit.ac.in/academics/schools/scope/faculty  
tripathybk@rediffmail.com, tripathyb@rediffmail.com, tripathyb@gmail.com

M. M. Kamruzzaman  
Assistant Professor Department of Computer and Information Science, Al Jouf University, Sakakah, Al Jouf, Saudi Arabia  
http://www.ju.edu.sa/index.php?id=1320&L=0  
m.m.kamruzzaman@gmail.com, kamruzzaman@qq.com, mmkamruzzaman@ju.edu.sa

Dr. Ali Taha Rashid Oqail  
Universiti Tenaga National, Malaysia  
Oqili83@yahoo.com

Ashwaq Talib Hashim  
Department of Control and Systems Engineering, University of Technology, Iraq  
ashwaqtalib@yahoo.com, 60102@uotechnology.edu.iq
Dr. Braham Deo Gupta  
Faculty, School of Mathematics and Allied Sciences, Jiwaji University, Gwalior, (M.P.) India  
https://www.researchgate.net/profile/Braham_Gupta  
brahamdeo.gupta@gmail.com

Dr. Ruzlaini Bt Ghoni  
Senior Lecturer, Department of Electrical/FTKEA Tati University College (TATIUC) Kolej Universiti TATI, Jalan Panchor, Telok Kalong, 24000 Kemaman, Terengganu, Malasya  
https://www.tatiuc.edu.my/assets/files/Dr.%20Ruzlaini%20-%20FTKEA.pdf  
ruzlaini@tatiuc.edu.my, ruzlainighoni@gmail.com

Dr. Sankalap Arora  
Assistant Professor, Department of Computer Science & Engineering, DAV University, Jalandhar, Punjab, INDIA  
https://davuniversity.org/faculty/dr-sankalap-arora-110.aspx  
sankalap.arora@gmail.com, sankalap10156@davuniversity.org

Dr. Vijay Khare  
Assistant Professor (SG) Electronic and Communication,Engineering Department Japyee Institute of Information Technology, Noida  
http://www.jiit.ac.in.academicsfacultyele  
vijay.khare@jiit.ac.in, vijay2512@gmail.com

Dr. SHANKAR R.  
Associate Professor, Department of Electronics & Communication Engineering (ECE), JEPPIAAR SRR Engineering College, Padur, Chennai, India  
http://www.jeppiaarsrr.in/jeppiaarsrr/ece.html  
dr.shankarr@hotmail.com

Dr. Ravindra Koggalage  
Head, Department of Electrical and Electronics Technology, University of Vocational Technology, Ratmalana, Sri Lanka  
http://univotec.ac.lk/index.php/academic-ftt/  
koggalage@yahoo.com, koggalage@univotec.ac.lk
List of reviewers


The Editorial team of the would like to thank the following reviewers for their work in refereeing manuscripts during 2018.

Raghu Vamsi Potukuchi
Kannadhasan Suriyan
Sridhar S
Ghossoon Mohammed Alsadoon
Gurjit Singh Bhathal
Mohd Muntjir
Ahmed Jabbar Abid
Davar Giveki
Shilpa Mehta
Dr Rajesh Ramachandran
Vijay Khare
Atanu Nag
Vinothkumar K
Ezhilarasi M
Miran Hikmat Baban
Evgeniy Lavrov
Nilu Singh
Ajith Kumar N
Mohamed Abdel Fatah Ashabrawy
Bhagwat Kakde
Dr Navaneethakrishnan R
Panneerselvam S
Md. Farukuzzaman Khan
Dr Kiran Sree Pokkakulri
Alireza Sedaghati
Arti Jain
Kiran Sree Pokkuluri
Shamik Tiwari
Hamdi Ercan
Abdallah Rhattoy
Appala Srinuvasu Muttipati
Parthiban Thanigachalam
Editorial

1. Annual Review Report

Research Articles

2. Close-up Macro and Micro Photogrammetry and Image Perspective: A Comparative Studio on different Lenses at work with Small and Medium Size Object
   Giorgio Verdiani, Paolo Formaglini, Filippo Giansanti; Stéphane Giraudieu 235-248

3. RLS Wiener Fixed-Point Smoother and Filter with Randomly Delayed or Uncertain Observations in Linear Discrete-Time Stochastic Descriptor Systems
   Seiichi Nakamori 249-266

4. On the Derivation and Analysis of a Highly Efficient Method for the Approximation of Quadratic Riccati Equations
   Joshua Sunday 267-281

5. Housekeeping Inspection and Inventory Analysis are the Primary Responses of Engineering and Logistics Operations in Hospitality Industry- An Intensive case study of Professional Research on Sheraton Gateway Hotel in Toronto Pearson International Airport
   Gazi Farok, Greg MacNeil 282-295

   Murtaza Mohiuddin Junaid Farooque 296-298

7. Designing a Customer Relationship Management System in Online Business
   Hamed Fazlollahtabar 299-312

   Jeremiah Ademola Balogun, Ngozi Chidozie Egejuru, Peter Adebayo Idowu 313-330

   Balogun Jeremiah Ademola, Aderounmu Temilade Egejuru Ngozi Chid, Idowu Peter Adebayo 331-346

10. A General Study on Langevin Equations of Arbitrary Order
    Elsayed M. Elsayed, S. Harikrishnan, K. Kanagarajan 347-360
Close-Up, Macro and Micro Photogrammetry and Image Perspective: A Comparative Studio on Different Lenses at Work with Small and Medium Size Objects

Giorgio Verdiani, Paolo Formaglini, Filippo Giansanti, Stéphane Giraudau

Architecture Photographic Laboratory, DIDALABS System, Dipartimento di Architettura, Florence University, Florence, Italy

giorgio.verdiani@unifi.it, paolo.formaglini@unifi.it, giansantifilippo@virgilio.it, s_giro@hotmail.com

Abstract:
The digital photogrammetry has renewed the approach to measurement for archaeologists, architects and many researchers, students, professionals in Cultural Heritage. Thus, most of the troubles coming from the more and more advanced software for photogrammetry processing came from purely photographic mistakes or poor knowledge about photographic tools. Here the focus will be on perspective and its influence in the result coming from medium and small size objects and finds. The study will present the results from the use of four different lenses for the same professional DSLR camera body: a Nikon D800e full frame 36 Megapixels, mounting the AF Micro-NIKKOR 60mm f/2.8D with 1:1 reproduction ratio (RT) (one of the best macro lens on the market), the Venus Laowa Micro 15mm f/4 with 1:1 RT (a quite economical super wide lens), a Nikkor Zoom 24-85mm AF with “macro” function (a classic in common set of lenses) and the classic Micro Nikkor 55mm F/2.8 with extension ring to reach the 1:1 RT (a piece of history of photography, the predecessor of the 60mm here in test). The full processing and procedure of matching the data will be presented to bring a useful contribution and reference for other scholars and operators.

Keywords: Photogrammetry; Close-up; 3d modeling; Digital photography; lenses;

Introduction

What does it mean macro photography? Macro photography is simply photographing small items, often insects and flowers, but also still life like jewelry and small household objects. Today it is also possible to consider the macro photography as an advanced survey instrument, which can become fundamental to museums or archaeological purposes. In the days of film, the answer to “what is macro photography” was a little stricter and required much more equipment. Shooting with a film camera, an image that captured something that was at least 1/10th of the original size on a piece of 35mm film was considered macro. Images that captured the object life-size, or at a 1:1 ratio, were considered micro. This can lead to some confusion between the micro and macro words, but commonly define macros all lenses that are able to focus very closely objects. Today, macro photography does not require special equipment; just think that some compact cameras have the macro function and are able to return quality images. It is unnecessary to add that a DSLR is the perfect choice because it’s possible choose and use of extremely high-quality lenses. The macro capabilities of a camera can be written in two different ways. The first is using ratio. A 1:1 macro image means that lens will capture a life-sized image of an object. A 1:2 ratio captures a small object at half of its original size, and vice versa. On DSLRs, macro capability is dependent on the lens, not the camera. Lenses with macro capabilities can get closer to a subject. A macro lens usually lists both a minimum focusing distance and a macro ratio in the technical specifications. A 1:1 is a great ratio for a DSLR lens, and many photographers consider anything less than 1:1 not a true macro. Macro lenses aren’t limited to just close-ups, either. They can also focus to infinity, meaning they can also snap non-macro photos too. More commonly, a camera or lenses macro capability is measured in the distance the object can be from the front of the lens. Precisely these factors led us to create a direct comparison between lenses using an archaeological find, 58mm high and 39mm wide. Furthermore, to make the test, we wanted more interesting to compare different types of lenses, mainly attracted from the exit on the market of a particular lens for its unique flow characteristics: the Laowa 15mm. This lens is very special: first, the wide focal is unused in macro photography, as preferred with smaller focal lenses; the 1:1 ratio reached as focus, indicates a good
quality of the lens construction; also has the possibility to shift the lens to correct any problems of perspective. Now we are going to see how we structured the workflow and the characteristics of the of selected lenses.

**Close-up, macro and micro photography**

The maximum magnification ratio is the minimum focus of the lens. 1: 1 means that the image is the same sensor size as it is. So, with a full frame sensor of 36x24 mm the width of the frame would be 36 mm. A 1: 2 the width of the frame filling the sensor would be 72 mm. The reproduction ratio is the reciprocal of the magnification that is, 1 divided by magnification.

We used full size of the frame sensor because they are standard, considering APS-C has different size depending on the manufacturer. However, in 1: 1 the image size is the same size of the sensor regardless of the size of the sensor. So, with APS-C image it would be more tightly framed. Whereas with average size, to 1: 1 more of the fractions would be in the frame.

How much of an advantage the maximum reproduction ratio or enlargement of a lens is, it all depends on what is the smallest thing is that you want to fill the frame. It has no real effect on a target to be used for portraiture, because you are usually some way off the minimum focusing distance, when taking a portrait photograph.

All lenses can be made to focus close with additional components such as extension tubes or close-up lenses, which increases the magnification. Thus, for example, with a 1:1 lens it is possible to lead something as large as 2: 1 (2x) using addons.

**MACRO RATIO SHOOTING**

![Macrophotography images](image)

Fig. 1 – Macro ratio shooting (Copyright: G. Verdiani, P. Formaglini, F. Giansanti, S. Girardeau)
Calibrated vs non-calibrated lenses

Nowadays there are several software applications (i.e., Photo modeler, Agisoft Lens, iWitness, etc.), mainly produced by the computer vision scientific community, that can automatically perform camera self-calibration. They also offer the possibility to work with several cameras and sensors to obtain dense point clouds or 3D models suitable for different fields of application.

In the wide panorama of consumer grade devices (including nowadays even smartphones or other similar mobile devices), the photogrammetric use though has not been easy since they could not supply high resolution still images and additionally their geometry is far away from the theoretical model of central projection due to their wide angle or fisheye lenses.

First, it is important to take into consideration distortions caused by the lens of camera; these can be divided into two types: geometrical distortions (radial and tangential) and chromatic aberrations. Radial distortion is a type of aberration causing straight lines to appear curved in the picture. It is caused by the misalignment of rays of light when crossing the group of lenses composing the entire lens. In the case where straight lines (referring to the middle of the picture) appear like concave curves, the radial distortion is so called “barrel distortion”. When those straight lines generate convex curves, the radial distortion is so called “pincushion distortion”. It is produced by the lens misalignments during construction phase, causing a shift of centers of the lenses from the main optical axis. Tangential distortion generates a compound deviation from radial and tangential component at the same time, being the latter orthogonal to the first component. However, in modern lenses this defect is negligible. Another lens defect is the Chromatic aberrations which causes red, blue, green, cyan, magenta or yellow halos, more noticeable in high contrast zones. They are caused by the different refraction index of the lens, according to the wavelength of light passing through the lens.

The testing grounds

Lenses used with a Nikon D800E - FF 36 Mpx:

Nikon AF-S 60mm Micro f/2.8G. One of the best lenses on market for macro-photography, at the same price of the 24-85mm tested. With 1:1 reproduction ratio.

Nikon 24-85mm f/2.8/4-AF-D-IF. Not a high-quality lens but a good one and not expensive for this kind of test. Versatile lens for any situation: landscape, portrait and macro-photography. With 1:2 reproduction ratio.

Nikon Micro-Nikkor 55mm f/2.8 + Nikon Extension ring PK-13 27.5mm. Old lens but good one, with a ratio of 1:2. With this extension ring the lens can reach the ratio of 1:1. Useful for macro-photography and for this kind of test.
Laowa 15mm f/4 Wide Angle 1:1 Macro Lens. This peculiar lens is very particular because it is at the same time wide angle and Macro with a RT up to 1:1, with also the possibility to shift. All these features in the same lens for a reasonable price in comparison with other better-known brands.

How to read the MTF chart

MTF (Modulation Transfer Function) is one of the measurements that evaluate a lens’ performance; it shows contrast reproducibility of the lens using characteristic spatial frequencies. Spatial frequencies indicate the number of lines per mm.

In the MTF the horizontal axis is in millimeters and shows the distance from the center of the image toward the edges, and contrast value (highest value is 1) is shown in the vertical axis, with fixed spatial frequencies of 10 lines/mm and 30 lines/mm.

The MTF chart for each lens is based on the value at the maximum aperture of the lens; the red line shows the spatial frequency of 10 lines/mm and the blue line, 30 lines/mm.

In the off-axis field, contrast reproducibility of the lens for sagittal direction and meridional direction varies with astigmatic affection. The path of 10 lines/mm indicates the contrast reproducibility of the lens (the higher and straighter is better). The higher and straighter the 30 lines/mm-path is, the higher the resolution of the lens. Note that the lens performance cannot be measured only with MTF chart. Softening or blurring of color also governs measurement.

Photographic set, lighting and object:

The object analyzed is a small artifact (58mm x 39mm), a fragment coming from a Saudi Arabian ancient vase.
Fig. 4 – The small artifact (Copyright: G. Verdiiani, P. Formaglini, F. Giansanti, S. Girardeau)

For a better framing, the object was put on the top of a tripod with a black backdrop behind.

Fig. 5 – Photographic set (Copyright: G. Author, P. Author, F. Author, S. Author)

The use of the tripod under the object was useful also for the possibility to keep very close the camera to the subject; the lamp, a Helios Biglamp 430 Ring 65W, with a folding stem, was the best choice because of its maneuverability and the opportunity to be close to the object without creating deep shadows.
Shooting on a tripod, we had the possibility to use a long-time exposure; the condition of shooting was the same for every lens: At f/11 S 1/50 ISO 400. The fact to use a f/11 was required to have a good depth of field. We decided to shoot at ISO 400 to not have an excessive time of exposure, and because of the high quality of the Nikon D800E, in facts the sensor of this camera produce almost noiseless images with this setting.

As we can see from the situation during the shooting the distance between the object and the focal plane was different for every lens. During the use of the Nikon AF-S 60mm Micro f/2.8G the distance was 24,4 cm; for the Nikon 24-85mm f/2.8/4-AF-D-IF the distance was 21,8 cm; instead for the Nikon Micro-Nikkor 55mm f/2.8 + Nikon Extension ring PK-13 27.5mm the distance was 26,5 cm and last the distance of the Laowa 15mm f/4 was just 12,1 cm, the shortest between all the lenses.
**First results**

Our first result has been to build, using software Agisoft PhotoScan, a series of four dense points clouds that would allow us to obtain a comparison between the results of the various lenses.

Among the four lenses choices, undoubtedly the best in terms of quality is to be the NIKON AF-S 60mm Micro f/2.8G. This aspect has led us to choose this lens as a reference model to compare the other lenses.

After performing the alignment of the pictures for every lens, operation that has not created problems in any of the four cases examined, we proceeded to the creation of dense and raw cloud in order to start the comparison. As we can see from the pictures above the next comparison about the points recognised for every lens highlights some differences on the correct distribution of points on the surface of the object.

As we expected the lens NIKON AF-S 60mm Micro f/2.8G gave a distribution of measured points on the surface homogeneous and widespread. This aspect is the first confirmation of the choice made about the reference lens.

The second lens that we used is a Laowa 15mm f/4 Wide Angle 1:1 Macro Lens. In that case the widespread is good but we have some parts, on bottom left corner, that has not detected points.

The dense clouds present a total of 9.723,781 points overall.

In the dense cloud of the Nikon 24-85mm f/2.8/4-AF-D-IF we can see a suboptimal distribution of the points and their absence in various parts of the object.

The dense clouds present a total of 9.723,781 points overall.

At the end we used Nikon Micro-Nikkor 55mm f/2.8 + Nikon Extension ring PK-13 27.5mm.
Analyzing the data extrapolated from the dense cloud of this lens we can see a good density of the points recognized and a widespread distribution over all the surface.

Fig. 9 – Points recognised for every lens (Copyright: G. Author, P. Author, F. Author, S. Author)

The total number of recognized points for this lens is 9,168,486.

Glancing at the final models can be an initial confirmation of the impressions we had from. While all models have a rather homogeneous although not totally collimating, the 85mm lens generates a cloud of points insufficient to accurately reconstruct the actual appearance of the object.

After these preliminary considerations we try to analyze the models obtained through the support of software “Raindrop Geomagic Qualify” for a more detailed analysis of the models.
Fig. 10 – 60mm model from Agisoft (Copyright: G. Verdi, P. Formaglini, F. Giansanti, S. Girardeau)

Fig. 11 – 55mm + ER27,5mm model from Agisoft (Copyright: G. Verdi, P. Formaglini, F. Giansanti, S. Girardeau)
The final benchmarks

After digital scanning and three-dimensional reconstruction of the object, the 3D models were compared to analyze the differences and the quality of the survey.
The first impression was positive, because all three lenses have produced results not too dissimilar from the reference model, that is the Nikon 60mm.

For the comparison of the 3D models, we used the Geomagic Qualify software that, using a color scale, makes a clear view of the differences between the models.

In detail, all three lenses have generated a good result but what has left us surprised was the quality achieved by 15mm. As it is possible to see from the graphs results obtained (fig. 14), the 15mm has produced the closest result to 60mm, with an average deviation of 0.037mm.

![Fig. 14 – Qualify Geomagic between 60mm and 15mm models (Copyright: G. Verdiiani, P. Formaglini, F. Giansanti, S. Girardeau)]

Good results have also been obtained with the 24–85mm that generated an average deviation equal to 0.04mm, that is very close to 15mm (fig. 15).
The 55mm, however, produced the worst results, with a gap of 0.062mm, probably due to the use of the extension ring (fig. 16).

It is important to note that the lower part of the model is very similar to the original and the main differences are in the parts in relief of the object.
Conclusions

The lenses we considered have shown different results that can induce some considerations about the quality in relation to accuracy in image reproduction, price and versatility:

**Nikon Micro-Nikkor 55mm f/2.8 + Nikon Extension ring PK-13 27.5mm**

Despite the contingent need of shooting conditions to bring the macro ratio to 1:2, this lens has proved to be the best from the point of view of homogeneity of the scan. Also, the result of points measured on the individual pictures of the scan revealed a large quantity of information. Due to an excessive prospective crushing if the limited range extension has been most lacking in raised dots.

**Nikon 24-85mm f/2.8-4-AF-D-IF**

Despite the zoom versatility that makes it suitable for all types of uses, from the calibration, this lens has managed to achieve good results both on flat surfaces and in those in relief. Unfortunately, the number of surface covered by aligned points is very low compared to the other lenses. As it can also be deduced from the decay of the MTF curves in the diagram.

**Nikon AF-S 60mm Micro f/2.8G**

This lens, with a focal length of 60mm, is the reference lens for macro photography with a focusing distance of 18.5cm from the focal plane. Thanks to the quality of the lenses, the Nikon 60mm is able to capture the finest details with no distortion and chromatic aberration. In our case, this lens has been used as a reference point for the other lenses, as it was assumed before the tests that could give better results due to its specific characteristics for the macro survey. The scan result has confirmed first impressions (due also to the MTF diagram), creating a cloud of points of quality and a subsequent mesh rich in detail, with around 14 million faces. It is noteworthy that in the scanning step, the results differ mainly for the different number of points of the cloud spread, which attests to the degree of correct points surveyed.

**Laowa 15mm f/4 Wide Angle 1:1 Macro Lens**

During this test we couldn’t use this lens at the ratio of 1:1 because of the excessive proximity of the lens to the object, just few millimeters from the top of the lens to the artifact, also the light couldn’t come in. Because of this situation the photos were taken at the ratio of 1:2 for all the lenses: in the case of the Laowa the object was just 2.4 cm from the lens. The decay of the MTF curves in the diagram of the Laowa is very similar to the 24-85mm, but about the aligned points it has behaved better than the other lens. This is an interesting lens and due to its versatility is to be appreciated more according to the results obtained, especially compared with other lenses.

This test has helped us to build a strong foundation to create a comparative database that we will push for a thorough comparison of professional lenses and entry level lenses.

**Funding Statement**

All the testing was hosted by the Architecture Photography Laboratory, DiDALabs System, the camera, lenses and accessories belongs to the Laboratory and/or to the Dipartimento di Architettura. The funding for this research comes from the DiDALabs System and/or from research on commission operated by the authors.
References


8. RAMÓN Y CAJAL, S., (1918). La microfotografía estereoscópica y biplanar del tejido nervioso.


RLS Wiener Fixed-Point Smoother and Filter with Randomly Delayed or Uncertain Observations in Linear Discrete-Time Stochastic Descriptor Systems

Seiichi Nakamori

Department of Technology, Faculty of Education, Kagoshima University, 1-20-6, Korimoto, Kagoshima, 890-0065 Japan

k6165161@kadai.jp

Abstract

The purpose of this paper is to design the recursive least-squares (RLS) Wiener fixed-point smoother and filter in linear discrete-time descriptor systems. The signal process is observed with additional observation noise. The observed value is randomly delayed by multiple sampling intervals or has the possibility of uncertainty that the observed value does not include the signal and contains the observation noise only. It is assumed that the probability of the observation delay and the probability that the observation does not contain the signal are known. The delayed or uncertain measurements are characterized by the Bernoulli random variables. The characteristic of this paper is that the RLS Wiener estimators are proposed from the randomly delayed, by multiple sampling intervals, or uncertain observations particularly for the descriptor systems in linear discrete-time stochastic systems.

Keywords: Discrete-time stochastic descriptor systems; RLS Wiener filter; RLS Wiener fixed-point smoother; randomly delayed observations; uncertain observations

1. Introduction

The descriptor systems have attracted attentions from the general expressions for physical models and structures in comparison with the conventional state-space models. The estimation problems for the descriptor systems have been studied ([1]-[8] and the references therein). Physical systems, e.g. a cart-pendulum system, electrical circuits [6], etc. are formulated as the descriptor systems. In [1], the recursive estimation algorithms for the filtering, prediction and fixed-lag smoothing estimates are presented, based on the innovation approach, for the descriptor systems with multiple packet dropouts and correlated noises. In [2], the information filter and predictor are proposed for the discrete-time descriptor systems with uncertain parameters. In addition to the nominal and robust estimation algorithms, the array algorithms are proposed. In [3], the robust Kalman type filter is proposed for the discrete-time descriptor systems with uncertainties in some matrices. In [4], according to the optimization technique, the optimum prediction estimate is updated in linear discrete-time descriptor systems. In [5], the robust predictor is presented for the discrete-time descriptor systems with bounded uncertainties. In [7], the problem of $H_{\infty}$ filtering for descriptor systems with strict linear matrix inequalities (LMIs) is investigated. The necessary and sufficient conditions for the solvability and the expression of the solution are obtained for both continuous-time and discrete-time descriptor systems. In [8], the paper studies on the delay-dependent robust $H_{\infty}$ filtering for uncertain discrete-time singular systems with the time-varying delay. Usually, the network data of control system are transmitted with delay and packet dropout from a sensor to a controller and also from a controller to an actuator [9]. In [10], for discrete-time stochastic linear systems with bounded random measurement delays and packet dropouts, the optimal filter, predictor and smoother are proposed based on the innovation approach. The observations are obtained in terms of (1c) in the paper. In [11], for the discrete-time stochastic systems with multiple packet dropouts, the optimal filter, predictor and smoother are proposed. Also, with observations multiply and randomly delayed, the recursive least-squares (RLS) Wiener fixed-point smoother and filter are proposed [12]. Under the condition that the uncertain observations are given, the RLS estimation algorithms are proposed, given the probability that the signal exists in the observation. The uncertain observation, if the signal exists, is characterized by using the independent Bernoulli random variable [13]. Estimation technique in [13] is extended to the case where the uncertain random variables are correlated [14]. In addition to the probability that the signal exists in the observation, the conditional probability is taken.
into account for the existence of the signal in the observation. In [15], the RLS Wiener fixed-point smoothing and filtering algorithms are proposed for the discrete-time stochastic system with uncertain observations. The probability that the signal exists in the observation and the conditional probability are required in the algorithms. With regard to multiple packet losses, related to the delayed observations, the optimal filter is devised in linear discrete-time stochastic systems over unreliable wireless sensor networks [16]. The technique is also extended to the extended Kalman filter in nonlinear discrete-time stochastic systems. In [17], the second order polynomial estimator is proposed in nonlinear systems with uncertain observations. The uncertainty in the observation equation is described by the Bernoulli random variables. In [18], the RLS Wiener fixed-point smoother and filter are proposed for the discrete-time stochastic systems with randomly delayed, by multiple sampling intervals, or uncertain observations. In [19], for the descriptor systems, based on the innovation theory, the RLS filter is proposed by using the covariance information of the state vector and the covariance information of the observation noise in linear discrete-time stochastic systems. Then the RLS Wiener type filter is presented for the descriptor systems.

In the packet dropout of the network systems, there might happen the case where the randomly delayed observed value does not include the signal. The packet dropout might not correspond to the certain observation including the signal, and also to the delayed observation. By the uncertain observation it means that there exists uncertainty if the observed value includes the signal or not [13]. In the packet dropout, the observation not including the signal data might further delay by one or more sampling intervals. As described above, the estimation problems for the descriptor systems have drawn great interests from the nature of the generalized state-space model. From this viewpoint, this paper, based on the estimation techniques both in [18] and [19], examines to design the RLS Wiener fixed-point smoother and filter from randomly delayed observed values, by multiple sampling intervals, or uncertain observations in linear discrete-time descriptor systems. The signal process is observed with additional observation noise. The observed value is randomly delayed by multiple sampling intervals or has the possibility of uncertainty that the observed value does not include the signal and contains the observation noise only. It is assumed that the probability of the random observation delay and the probability that the observation does not contain the signal are known as a priori information. The randomly delayed or uncertain measurements are characterized by the Bernoulli random variables. The characteristic of this paper is that the RLS Wiener estimators are proposed from the randomly delayed, by multiple sampling intervals, or uncertain observations particularly for the descriptor systems in linear discrete-time stochastic systems.

A numerical simulation example, in section 5, shows the estimation characteristics of the proposed fixed-point smoother and filter with the randomly delayed, by multiple sampling intervals, or uncertain observations in linear discrete-time descriptor systems.

2. Least-squares smoothing problem with delayed or uncertain observations for descriptor systems

In linear discrete-time descriptor systems, the state and observation equations are described by [1]

$$\Xi S(k+1) = FS(k) + \Lambda w(k), E[w(k)]w^T(s)] = Q\delta_k(k-s),$$

$$\bar{y}(k) = CS(k) + v(k), E[v(k)v^T(s)] = R\delta_k(k-s).$$

(1)

Here, $S(k)$ represents an $n$-dimensional descriptor vector, $w(k)$ a $q$-dimensional input noise and $\bar{y}(k)$ an $m$-dimensional certain measurement without including delays or uncertain signals. $\Xi$, $F$, $\Lambda$ and $C$ are matrices with the dimensions $n \times n$, $n \times n$, $n \times q$, $m \times n$ respectively. In the descriptor systems, $\Xi$ is the singular matrix, i.e. \(\text{rank}(\Xi) < n\). In terms of orthogonal matrices $U$ and $V$, the singular value decomposition (SVD) of $\Xi$ is written as follows.
\[
\Xi = UDV^T, \quad D = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}, \quad U^T = U^{-1}, \quad V^T = V^{-1},
\]
\[
\Lambda = \text{diag}(\mu_1, \mu_2, \ldots, \mu_l), \quad \mu_i > 0, i = 1, 2, \ldots, l, \Delta > 0
\]  

(2)

From the relationships

\[
U^T F V = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}, \quad U^T \Lambda = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix}, \quad CV = [C_1 \quad C_2], \quad S(k) = V \begin{bmatrix} \tilde{S}_1(k) \\ \tilde{S}_2(k) \end{bmatrix},
\]

the state equation in (1) is described by

\[
\begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{S}_1(k+1) \\ \tilde{S}_2(k+1) \end{bmatrix} = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \begin{bmatrix} \tilde{S}_1(k) \\ \tilde{S}_2(k) \end{bmatrix} + \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix} w(k).
\]

(3)

From (3), it follows that

\[
\tilde{S}_1(k+1) = A \tilde{S}_1(k) + Bw(k), \quad A = \Delta^{-1}(F_{11} + F_{12} \Gamma_1), \quad B = \Delta^{-1}(F_{12} \Gamma_2 + \Lambda_1) > 0
\]

\[
\tilde{y}(k) = H \tilde{S}_1(k) + \tilde{\nu}(k), \quad H = C_1 + C_2 \Gamma_1, \quad \tilde{\nu}(k) = C_2 \Gamma_2 w(k) + v(k),
\]

\[
E[\tilde{y}(k)\tilde{y}^T(s)] = R \delta_k (k-s), \quad R = C_2 \Gamma_2 Q C^*_2 + R,
\]

\[
\tilde{S}_2(k) = \Gamma_1 \tilde{S}_1(k) + \Gamma_2 w(k), \quad \Gamma_1 = -F_{21}^{-1}F_{22} \Gamma_2 = -F_{22}^{-1} \Lambda_2, \quad F_{22} > 0.
\]

(4)

For notational conveniences, let us put \( \tilde{S}_1(k) \) as \( x(k) = \tilde{S}_1(k) \) and \( A \) as \( \Phi = A \). Then the state equation for the state vector \( x(k) \) and its observation equation are given by

\[
x(k+1) = \Phi x(k) + Bw(k), \quad E[w(k)w^T(s)] = Q \delta_k (k-s),
\]

\[
\tilde{y}(k) = z(k) + \tilde{\nu}(k), \quad z(k) = Hx(k), \quad E[\tilde{y}(k)\tilde{y}^T(s)] = R \delta_k (k-s).
\]

(5)

For the discrete-time systems with measurement delays or uncertain observations, let an m-dimensional observation equation be described as

\[
\begin{align*}
\gamma(k) &= \gamma_{00}(k) \tilde{y}(k) + \gamma_{11}(k) \tilde{y}(k-1) + \cdots + \gamma_{11}(k) \tilde{y}(k-N) + \gamma_{00}(k) \tilde{y}(k-1) + \cdots \\
&\quad + \gamma_{00}(k) \tilde{y}(k-N), \\
\tilde{y}(k) &= z(k) + \tilde{\nu}(k), \quad z(k) = Hx(k),
\end{align*}
\]

\[
E[\gamma_{ij}(k)] = \rho_{00}(k), E[\gamma_{11}(k)] = \rho_{11}(k), E[\gamma_{21}(k)] = \rho_{21}(k), \ldots, E[\gamma_{R1}(k)] = \rho_{R1}(k),
\]

\[
E[\gamma_{00}(k)] = \rho_{00}(k), E[\gamma_{10}(k)] = \rho_{10}(k), E[\gamma_{20}(k)] = \rho_{20}(k), \ldots, E[\gamma_{R0}(k)] = \rho_{R0}(k).
\]

(6)

Let us assume that the observation at each time \( k > 1 \) can either be delayed by sampling intervals \( j, \) \( 1 \leq j \leq N \), with known probabilities or consists of delayed measurements, which do not contain signal data. \( \{\gamma_{ij}(k), 0 \leq i \leq N, j = 0, 1; \ k > 1\} \) represent a sequence of Bernoulli random variables (binary switching sequence...
\[
\mathcal{F}_i(k) = \begin{bmatrix}
\gamma_{i0}(k) & I_{new} & \gamma_{i1}(k) & I_{new} & \gamma_{i2}(k) & I_{new} & \ldots & \gamma_{iN}(k) & I_{new}
\end{bmatrix},
\]

\[
\mathcal{F}_0(k) = \begin{bmatrix}
\gamma_{00}(k) & I_{new} & \gamma_{01}(k) & I_{new} & \gamma_{02}(k) & I_{new} & \ldots & \gamma_{0N}(k) & I_{new}
\end{bmatrix},
\]

taking the values 0 or 1 with
\[
\mathcal{F}_i(k) = \begin{bmatrix}
\bar{\gamma}(k) & \bar{\gamma}(k-1) & \ldots & \bar{\gamma}(k-N)
\end{bmatrix}^T,
\]
\[
\mathcal{F}_0(k) = \begin{bmatrix}
\bar{\gamma}(k) & \bar{\gamma}(k-1) & \ldots & \bar{\gamma}(k-N)
\end{bmatrix}^T.
\]

\[
P[\gamma_j(k) = 1] = p_j(k), 0 \leq i \leq \bar{N}, j = 0, 1).\]

By introducing the notations
\[
(7)
\]
from (6), we obtain
\[
y(k) = \mathcal{F}_i(k) \bar{\gamma}(k) + \mathcal{F}_0(k) \bar{v}(k).
\]

\[(8)\]
\[
\mathcal{F}_i(k) \quad \text{corresponds to the Bernoulli random variables for the measurement delays and} \quad \mathcal{F}_0(k) \quad \text{for the observations, which consist of only observation noise data. Let} \quad E_p[.] \quad \text{represents the expectation with respect to the random variables} \quad \{\gamma(k), k \geq 1\}. \quad \text{The Bernoulli random variables satisfy} \quad E_p[\gamma_j(k)] = p_j(k)I_{new},
\]

\[
E_p[\gamma_j^2(k)] = p_j(k)I_{new}, 0 \leq i \leq \bar{N}, j = 0, 1. \quad \text{It is found that the auto-covariance function} \quad K_{\bar{v}}(k, s) \quad \text{of} \quad \bar{v}(k) \quad \text{is given by}
\]

\[
K_{\bar{v}}(k, s) = \begin{bmatrix}
\bar{C}(k)\bar{D}^T(s), 0 \leq s \leq k, \\
\bar{D}(k)\bar{C}^T(s), 0 \leq k \leq s,
\end{bmatrix}
\]

\[(9)\]
\[
\bar{C}(k) = \Phi_{\bar{v}}^k, \bar{D}^T(s) = \Phi_{\bar{v}}^sK_{\bar{v}}(s, s). \quad \text{Here, the transition matrix} \quad \Phi_{\bar{v}} \quad \text{and the variance} \quad K_{\bar{v}}(s, s) \quad \text{of} \quad \bar{v}(k) \quad \text{are given as follows.}
\]

\[
\Phi_{\bar{v}} = \begin{bmatrix}
0 & 0 & \ldots & 0 & 0 \\
I_{new} & 0 & \ldots & 0 & 0 \\
\vdots & \ddots & \ddots & \vdots & \vdots \\
0 & \ldots & 0 & I_{new} & 0
\end{bmatrix}, K_{\bar{v}}(s, s) = \begin{bmatrix}
\bar{R} & 0 & \ldots & 0 \\
0 & \bar{R} & \ldots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \ldots & 0 & \bar{R}
\end{bmatrix}
\]

\[(10)\]

By introducing
\[
\bar{H} = \begin{bmatrix}
H & \ldots & 0 & 0 \\
0 & \ldots & H & 0 \\
0 & \ldots & 0 & H
\end{bmatrix}, \quad \bar{x}(k) = \begin{bmatrix}
x(k) \\
x(k-1) \\
\vdots \\
x(k-N)
\end{bmatrix},
\]

\[(11)\]

from (6) and (7), the observation equation (8) is rewritten as
\[
y(k) = \mathcal{F}_i(k)\bar{H}\bar{x}(k) + \mathcal{F}_i(k)\bar{v}(k) + \mathcal{F}_0(k)\bar{v}(k),
\]

\[(12)\]
\[
\begin{bmatrix}
z(k) \\
z(k-1) \\
\vdots \\
z(k-N)
\end{bmatrix} = 
\begin{bmatrix}
Hx(k) \\
Hx(k-1) \\
\vdots \\
Hx(k-N)
\end{bmatrix} = \bar{H}_x(k).
\] (13)

Let \( K_x(k, s) \) denote the auto-covariance function of the state vector \( x(k) \) in wide-sense stationary stochastic systems [20], and let \( K_x(k, s) \) be expressed by

\[
K_x(k, s) = \begin{cases} 
A(k)B^T(s), & 0 \leq s \leq k, \\
B(s)A^T(k), & 0 \leq k \leq s,
\end{cases}
\] (14)

\[
A(k) = \Phi^k, \quad B^T(s) = \Phi^{-s}K_x(s, s). \quad \text{Here, } \Phi \text{ represents the transition matrix of } x(k). \quad \text{From}
\]

\[
\begin{bmatrix}
x(k+1) \\
x(k) \\
\vdots \\
x(k-N+2) \\
x(k-N+1)
\end{bmatrix} = 
\begin{bmatrix}
\Phi & 0 & \cdots & 0 & 0 \\
I_{n} & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cdots & I_{n} & 0
\end{bmatrix} 
\begin{bmatrix}
x(k) \\
x(k-1) \\
\vdots \\
x(k-N+1) \\
x(k-N)
\end{bmatrix} + 
\begin{bmatrix}
w(k) \\
0 \\
\vdots \\
0 \\
0
\end{bmatrix},
\] (15)

The system matrix \( \Phi \) for the state vector \( \tau(k) \) is given by

\[
\Phi = 
\begin{bmatrix}
\Phi & 0 & \cdots & 0 & 0 \\
I_{n} & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 \\
0 & 0 & \cdots & I_{n} & 0
\end{bmatrix}
\] (16)

Let \( K_x(k, s) \) represent the auto-covariance function of \( \tau(k) \). Then \( K_x(k, s) \) is given by

\[
K_{\tau}(k, s) = \begin{cases} 
\tilde{A}(k)B^T(s), & 0 \leq s \leq k, \\
\tilde{B}(k)\tilde{A}^T(s), & 0 \leq k \leq s,
\end{cases}
\] (17)

\[
\tilde{A}(k) = \Phi^k, \quad B^T(s) = \Phi^{-s}K_{\tau}(s, s). \quad \text{Here, } K_{\tau}(s, s) = K_{\tau}(0) \text{ is described as}
\]
Let the fixed-point smoothing estimate \( \hat{x}(k, L) \) of \( \bar{x}(k) \) at the fixed point \( k \) be given by

\[
\hat{x}(k, L) = \sum_{i=1}^{L} h(k, i, L) y(i)
\]  

(19)
in terms of the observed values \( \{y(i), 1 \leq i \leq L\} \). In (19), \( h(k, i, L) \) is a time-varying impulse response function. We consider the estimation problem, which minimizes the mean-square value (MSV)

\[
J = E[\|\bar{x}(k) - \hat{x}(k, L)\|^2]
\]  

(20)
of the fixed-point smoothing error. Based on an orthogonal projection lemma [20],

\[
\bar{x}(k) - \sum_{i=1}^{L} h(k, i, L) y(i) \perp y(s), 1 \leq s \leq L,
\]  

(21)
the optimal impulse response function satisfies the Wiener-Hopf equation

\[
E[\bar{x}(k) y^T(s)] = \sum_{i=1}^{L} h(k, i, L) E[y(i) y^T(s)].
\]  

(22)
Here \( \perp \) denotes the notation of the orthogonality. By introducing

\[
\bar{P}_k(k) = \begin{bmatrix} p_{01}(k)I_{\text{msm}} & p_{11}(k)I_{\text{msm}} & \cdots & p_{R-11}(k)I_{\text{msm}} & p_{R1}(k)I_{\text{msm}} \end{bmatrix},
\]

from (7) and (12), the left hand side of (22) is developed as

\[
E[\bar{x}(k) y^T(s)] = E[\bar{P}_k(k) \bar{\Omega}(s) \bar{\Omega}(s)^T + \bar{P}_0(s) \bar{v}(s) + \bar{P}_0(s) \bar{v}(s)^T]
\]

(23)
\[= E_p[\gamma_1(i)\bar{H}\bar{K}_\tau(i,s)\bar{H}^T\gamma_1^T(s)] + E_p[\gamma_2(i)\bar{K}_\psi(i,s)\gamma_2^T(s)], \]

\[\gamma_2(s) = \gamma_0(s) + \gamma_1(s). \quad (24)\]

Substituting (23) and (24) into (22), we have

\[K_\tau(k,s)H^T\gamma_1^T(s) = \sum_{i=1}^{L} h(k,i,L)\{E_p[\gamma_1(i)\bar{H}\bar{K}_\tau(i,s)\bar{H}^T\gamma_1^T(s)] + E_p[\gamma_2(i)\bar{K}_\psi(i,s)\gamma_2^T(s)]\}. \quad (25)\]

From the stochastic property of \(\gamma_1(\cdot)\) and \(\gamma_2(\cdot)\), (25) is rewritten as

\[K_\tau(k,s)H^T\gamma_1^T(s) = h(k,s,L)\{E_p[\gamma_1(s)\bar{H}\bar{K}_\tau(s,s)\bar{H}^T\gamma_1^T(s)] + E_p[\gamma_2(s)\bar{K}_\psi(s,s)\gamma_2^T(s)]\] 

\[-E_p[\gamma_1(s)\bar{H}\bar{K}_\tau(s,s)\bar{H}^T\gamma_1^T(s)] - E_p[\gamma_2(s)\bar{K}_\psi(s,s)\gamma_2^T(s)]\} + \sum_{i=1}^{L} h(k,i,L)\{E_p[\gamma_1(i)\bar{H}\bar{K}_\tau(i,s)\bar{H}^T\gamma_1^T(s)] + E_p[\gamma_2(i)\bar{K}_\psi(i,s)\gamma_2^T(s)]\}. \quad (26)\]

Rearranging (26) and introducing \(R(s)\), we obtain the equation for the optimal impulse response function \(h(k,s,L)\) as

\[h(k,s,L)\bar{R}(s) = K_\tau(k,s)H^T\gamma_1^T(s) \]

\[= \sum_{i=1}^{L} h(k,i,L)\{\gamma_1(i)\bar{H}\bar{K}_\tau(i,s)\bar{H}^T\gamma_1^T(s) + \gamma_2(i)\bar{K}_\psi(i,s)\gamma_2^T(s)\}, \quad (27)\]

\[\bar{R}(s) = E_p[\gamma_1(s)\bar{H}\bar{K}_\tau(s,s)\bar{H}^T\gamma_1^T(s)] + E_p[\gamma_2(s)\bar{K}_\psi(s,s)\gamma_2^T(s)] \]

\[-E_p[\gamma_1(s)\bar{H}\bar{K}_\tau(s,s)\bar{H}^T\gamma_1^T(s)] - E_p[\gamma_2(s)\bar{K}_\psi(s,s)\gamma_2^T(s)]\}. \quad (28)\]

3. RLS Wiener estimation algorithms with delayed or uncertain measurements for descriptor systems

Under the problem formulation in section 2 on the linear least-squares estimation for the descriptor systems with the randomly delayed, by multiple sampling intervals, or uncertain observations, Theorem 1 presents the RLS Wiener fixed-point smoothing and filtering algorithms of the descriptor vector \(S(k)\).

**Theorem 1**

Based on the optimal estimation problems in section 2, the RLS Wiener algorithms for the fixed-point smoothing and filtering estimates of the descriptor vector \(S(k)\) consist of (29)-(47) for the descriptor systems with randomly delayed, by multiple sampling intervals, or uncertain observations in linear discrete-time stochastic systems.

Fixed-point smoothing estimate of the descriptor vector \(S(k) = V\begin{bmatrix} \hat{S}_1(k) \\ \hat{S}_2(k) \end{bmatrix} \) : \(\hat{S}(k,L)\)

\[\hat{S}(k,L) = V\begin{bmatrix} \hat{S}_1(k,L) \\ \hat{S}_2(k,L) \end{bmatrix} \quad (29)\]

Fixed-point smoothing estimate of \(\hat{S}_1(k)\) at the fixed point \(k : \hat{S}_1(k,L)\)
\[
\hat{S}_1(k, L) = [I_{l \times l} \ 0_{l \times N}] \hat{x}(k, L)
\]  

(30)

Fixed-point smoothing estimate of \( \bar{S}_2(k) : \hat{S}_2(k, L) \)

\[
\hat{S}_2(k, L) = \Gamma_1 \hat{S}_1(k, L),
\]

\[
\Gamma_1 = -F_{22}^{-1} F_{21}, \Xi = UDV^T, \ D = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}, U^T = U^{-1}, V^T = V^{-1},
\]

\[
\Delta = \text{diag}(\mu_1, \mu_2, ..., \mu_l), \mu_i > 0, i = 1, 2, ..., l, \Delta > 0, U^T F V = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}
\]  

(31)

Filtering estimate of the descriptor vector \( S(k) = V \begin{bmatrix} \hat{S}_1(k) \\ \hat{S}_2(k) \end{bmatrix} : \hat{S}(k, L) \)

\[
\hat{S}(k, L) = V \begin{bmatrix} \hat{S}_1(k, k) \\ \hat{S}_2(k, k) \end{bmatrix}
\]  

(32)

Filtering estimate of \( \bar{S}_1(k) : \hat{\bar{S}}_1(k, k) \)

\[
\hat{\bar{S}}_1(k, k) = [I_{l \times l} \ 0_{l \times N}] \hat{x}(k, k)
\]  

(33)

Filtering estimate of \( \bar{S}_2(k) : \hat{\bar{S}}_2(k, k) \)

\[
\hat{\bar{S}}_2(k, k) = \Gamma_1 \hat{\bar{S}}_1(k, k)
\]  

(34)

Fixed-point smoothing estimate of \( \bar{x}(k) \) at the fixed point \( \hat{k} : \hat{\bar{x}}(k, L) \)

\[
\hat{\bar{x}}(k, L) = \hat{\bar{x}}(k, L-1) + h(k, L, L)(y(L) - \bar{p}_1(L)\bar{H} \bar{\Phi} \hat{\bar{x}}(L-1, L-1) - \bar{p}_2(L)\Phi \hat{v}(L-1, L-1)),
\]

(35)

Smooother gain: \( h(k, L, L) \)

\[
h(k, L, L) = [K_x(k, k)(\bar{\Phi}^T)^{-k} \bar{H}^T \bar{p}_1^T(L) - q_1(k, L-1)\bar{\Phi}^T \bar{H}^T \bar{p}_1^T(L) - q_2(k, L-1)\Phi \bar{p}_2^T(L)]
\]

\[
\times [\hat{\bar{R}}(L) + \bar{p}_1(L)\bar{H} \bar{K}_x(L, L) - \bar{p}_1(L)\bar{H} \bar{\Phi} \bar{S}_1(L-1, L-1)\bar{\Phi}^T - \bar{p}_2(L)\Phi \bar{S}_12(L-1, L-1)\bar{\Phi}^T] \bar{H}^T \bar{p}_1^T(L)
\]

\[
+ [\bar{p}_2(L)\Phi \bar{S}_12(L-1, L-1)\bar{\Phi}^T - \bar{p}_2(L)\Phi \bar{S}_22(L-1, L-1)\Phi \bar{p}_2^T(L)] \bar{H}^T \bar{p}_2^T(L)^{-1}
\]  

(36)
\[ q_1(k, L) = q_1(k, L-1)\Phi^T + h(k, L, L)(\bar{p}_1(L)H\bar{K}_x(L, L) - \bar{p}_1(L)H\bar{S}_{11}(L-1)\Phi^T \\
- \bar{p}_1(L)\Phi \bar{S}_{21}(L-1)\Phi^T]. \] (37)

\[ q_1(k, k) = S_{11}(k) \]

\[ q_2(k, L) = q_2(k, L-1)\Phi^T + h(k, L, L)(\bar{p}_2(L)K_y(L, L) - \bar{p}_1(L)H\bar{S}_{12}(L-1)\Phi^T \\
- \bar{p}_2(L)\Phi \bar{S}_{22}(L-1)\Phi^T], \] (38)

\[ q_2(k, k) = S_{12}(k) \]

Filtering estimate of \( \hat{x}(L) : \hat{x}(L, L) \)

\[ \hat{x}(L, L) = \bar{K}\hat{x}(L-1, L-1) + G_1(L, L)(y(L) - \bar{p}_1(L)H\bar{\hat{x}}(L-1, L-1) \\
- \bar{p}_2(L)\Phi \hat{v}(L-1, L-1)), \] (39)

\[ \hat{x}(0,0) = 0 \]

Filtering estimate of \( \hat{v}(L) : \hat{v}(L, L) \)

\[ \hat{v}(L, L) = \Phi \hat{v}(L-1, L-1) + G_2(L, L)(y(L) - \bar{p}_1(L)H\bar{\hat{x}}(L-1, L-1) \\
- \bar{p}_2(L)\Phi \hat{v}(L-1, L-1)), \] (40)

\[ \hat{v}(0,0) = 0 \]

Auto-variance function of \( \hat{x}(L, L) : S_{11}(L) = E[\hat{x}(L, L)\hat{x}^T(L, L)] \)

\[ S_{11}(L) = \bar{S}S_{11}(L-1)\Phi^T + G_1(L, L)(\bar{p}_1(L)H\bar{K}_x(L, L) - \bar{p}_1(L)H\bar{S}_{11}(L-1)\Phi^T \\
- \bar{p}_2(L)\Phi \bar{S}_{21}(L-1)\Phi^T], \] (41)

\[ S_{11}(0) = 0 \]

Cross-variance function of \( \hat{x}(L, L) \) with \( \hat{v}(L, L) : S_{12}(L) = E[\hat{x}(L, L)\hat{v}^T(L, L)] \)

\[ S_{12}(L) = \bar{S}S_{12}(L-1)\Phi^T + G_1(L, L)(\bar{p}_2(L)K_y(L, L) - \bar{p}_1(L)H\bar{S}_{12}(L-1)\Phi^T \\
- \bar{p}_2(L)\Phi \bar{S}_{22}(L-1)\Phi^T], \] (42)

\[ S_{12}(0) = 0 \]

\[ S_{21}(L) = \Phi \bar{S}_{21}(L-1)\Phi^T + G_2(L, L)(\bar{p}_1(L)H\bar{K}_x(L, L) - \bar{p}_1(L)H\bar{S}_{11}(L-1)\Phi^T \\
- \bar{p}_2(L)\Phi \bar{S}_{21}(L-1)\Phi^T], \] (43)

\[ S_{21}(0) = 0, S_{21}(L) = S_{21}^T(L) \]

Auto-variance function of \( \hat{v}(L, L) : S_{22}(L) = E[\hat{v}(L, L)\hat{v}^T(L, L)] \)

\[ S_{22}(L) = \Phi \bar{S}_{22}(L-1)\Phi^T + G_2(L, L)(\bar{p}_2(L)K_y(L, L) - \bar{p}_1(L)H\bar{S}_{12}(L-1)\Phi^T \\
- \bar{p}_2(L)\Phi \bar{S}_{22}(L-1)\Phi^T], \] (44)

\[ S_{22}(0) = 0 \]
\[ G_1(L, L) = [K_x(L, L)\Phi_T(L) - \Phi S_{11}(L-1)\Phi_T(L) - \Phi S_{12}(L-1)\Phi_T p_2(L)] \]
\[ \times \{ [\Phi v(L) - \Phi S_{11}(L-1)\Phi_T p_2(L)] - \Phi S_{12}(L-1)\Phi_T p_2(L) \} \]
\[ + [\Phi v(L) - \Phi S_{11}(L-1)\Phi_T p_2(L)] \Phi_T p_2(L)^{-1} \]
\[ G_2(L, L) = [K_x(L, L)p_1(L)\Phi_T(L) - \Phi S_{11}(L-1)\Phi_T p_1(L) - \Phi S_{12}(L-1)\Phi_T p_1(L)] \]
\[ \times \{ [\Phi v(L) - \Phi S_{11}(L-1)\Phi_T p_1(L)] - \Phi S_{12}(L-1)\Phi_T p_1(L) \} \]
\[ + [\Phi v(L) - \Phi S_{11}(L-1)\Phi_T p_1(L)] \Phi_T p_1(L)^{-1} \]
\[ R(L) = E_x[\Phi v(L) + [\Phi v(L)] - E_x[p_1(L)K_v(L, L)p_1(L)] \]
\[ - E_x[p_1(L)K_v(L, L)] \Phi_T p_1(L)] - E_x[p_1(L)]K_v(L, L) E_x[p_1(L)] \]

Theorem 1 is proved by referring to the RLS Wiener filter and fixed-point smoother for the systems with randomly delayed, by multiple sampling intervals, or uncertain observations [18] and the RLS Wiener filter for the descriptor systems [19] in linear discrete-time stochastic systems.

As the condition for the asymptotic stability of the filtering equation in Theorem 1 for \( \hat{x}(L, L) \), it is necessary that all the eigenvalues of \( \Phi - G_1(L, L)p_1(L)\Phi \) lie inside the unit circle. In addition, for the stability of the estimation algorithms, from (36), (45) and (46), the following matrix must be positive definite.

\[ \bar{R}(L) + [\Phi v(L) - \Phi S_{11}(L-1)\Phi_T p_1(L)] - \Phi S_{12}(L-1)\Phi_T p_1(L) \]
\[ + [\Phi v(L) - \Phi S_{11}(L-1)\Phi_T p_1(L)] \Phi_T p_1(L)^{-1} \]

4. A numerical simulation example

Let a scalar observation equation be given by

\[ y(k) = CS(k) + v(k), E[v(k)v(s)] = R\delta_k(k-s), \]

for the descriptor system

\[ \Xi S(k+1) = FS(k) + \Lambda w(k), E[w(k)w(s)] = Q\delta_k(k-s), \]

\[ S(k) = \begin{bmatrix} S_1(k) \\ S_2(k) \end{bmatrix}, \Xi = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, F = \begin{bmatrix} 0 & 1 \\ -0.8 & -0.1 & 0.5 \end{bmatrix}, \Lambda = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}, Q = 0.5^2, C = [0.5 \ 0.9 \ 0.6], \]

in linear discrete-time stochastic systems. Here, \( S(k) \) is the descriptor vector, \( w(k) \) denotes the input noise and \( y(k) \) is the certain measurement without considering measurement delays or uncertain signals. Here, \( \Xi \) is the singular matrix, i.e. \( rank(\Xi) = 2 \leq 3 \). With orthogonal matrices \( U \) and \( V \), the SVD of \( \Xi \) is expressed by

\[ \Xi = UDV^T, D = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}, U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \Delta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \]

From the relationships
\[ U^T F V = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}, \quad U^T \Lambda = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix}, \quad CV = [C_1 \quad C_2], \quad S(k) = \begin{bmatrix} S_1(k) \\ S_2(k) \\ S_3(k) \end{bmatrix} = V \begin{bmatrix} S_1(k) \\ S_2(k) \end{bmatrix}, \]

the state equation in (49) and the observation equation in (48),

\[ \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \bar{S}_1(k+1) = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \bar{S}_1(k) + \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix} w(k), \]

\[ \bar{y}(k) = C \bar{S}_1(k) + v(k), E[v(k)v(s)] = R \delta_k(k-s), \]

\[ F_{11} = \begin{bmatrix} 0 & 1 \\ -0.8 & -0.1 \end{bmatrix}, F_{12} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}, F_{21} = [0 \quad 0.5], F_{22} = 1.5, \]

\[ \Lambda_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}, \Lambda_2 = 0.2, \]

\[ C_1 = [0.5 \quad 0.9], C_2 = 0.6. \]

are transformed into

\[ \bar{S}_1(k+1) = A \bar{S}_1(k) + B w(k), A = \Delta^{-1}(F_{11} + F_{12} \Gamma_1), B = \Delta^{-1}(F_{12} \Gamma_2 + \Lambda_2), \]

\[ \bar{y}(k) = H \bar{S}_1(k) + \bar{v}(k), H = C_1 + C_2 \Gamma_1, \bar{v}(k) = C_2 \Gamma_2 w(k) + v(k), \]

\[ E[\bar{v}(k)\bar{v}(s)] = R \delta_k(k-s), \]

\[ B = C_1 \Gamma_2 \Gamma_1^T C_1^T + R, \]

\[ \bar{S}_2(k) = \Gamma_1 \bar{S}_1(k) + \Gamma_2 w(k), \Gamma_1 = -F_{22}^{-1} F_{21}, \Gamma_2 = -F_{22}^{-1} \Lambda_2, F_{22} > 0. \]

By putting \( \bar{S}_1(k) \) as \( x(k) = \bar{S}_1(k) \) and \( \Lambda \) as \( \Phi = A \), the state equation for the state vector \( x(k) \) and its observation equation, without measurement delays or including uncertain observations, are as follows.

\[ x(k+1) = \Phi x(k) + B w(k), E[w(k)w(s)] = Q \delta_k(k-s), \]

\[ \bar{y}(k) = z(k) + \bar{v}(k), z(k) = H x(k), E[\bar{v}(k)\bar{v}(s)] = R \delta_k(k-s). \]

Now, let us consider the observation equation in the case of \( \bar{N} = 2 \) in (6).

\[ \gamma(k) = \bar{y}_1(k) \bar{v}(k) + \bar{y}_0(k) \bar{v}(k) = \bar{y}_1(k) \bar{H} \bar{x}(k) + \bar{y}_2(k) \bar{v}(k), \]

\[ \bar{y}_1(k) = [\gamma_{01}(k) \quad \gamma_{11}(k) \quad \gamma_{21}(k)], \bar{y}_0(k) = [\gamma_{00}(k) \quad \gamma_{10}(k) \quad \gamma_{20}(k)]. \]

\[ \bar{y}_1(k) = [\gamma_{01}(k) + \gamma_{10}(k) \quad \gamma_{11}(k) + \gamma_{20}(k) \quad \gamma_{21}(k)], \bar{y}_0(k) = [\gamma_{00}(k) \quad \gamma_{10}(k) \quad \gamma_{20}(k)]. \]

\[ \bar{y}_1(k) = [p_{01}(k) \quad p_{11}(k) \quad p_{21}(k)], \bar{y}_0(k) = [p_{00}(k) \quad p_{10}(k) \quad p_{20}(k)]. \]

\[ \bar{y}(k) = [\bar{v}(k) \quad \bar{v}(k-1) \quad \bar{v}(k-2)]^T, \]

\[ \bar{v}(k) = [\bar{v}(k) \quad \bar{v}(k-1) \quad \bar{v}(k-2)]^T, \]

\[ \bar{z}(k) = \bar{z}(k-1) = [H x(k)], \bar{z}(k-2) = [H x(k-1)], \bar{z}(k-2) = [H x(k-2)]. \]

\[ \tilde{H} = \begin{bmatrix} H & 0 & 0 \\ 0 & H & 0 \\ 0 & 0 & H \end{bmatrix} = \begin{bmatrix} 0.5 & 0.7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.7 \end{bmatrix}, H = [0.5 \quad 0.7]. \]
The values of the probabilities for the Bernoulli random variables, $\gamma_i(k), i, j = 0, 1, 2$, satisfy

$$
\begin{align*}
\Pr[\gamma_0(k)] &= \Pr[\gamma_0^2(k)] = p_{00}, \\
\Pr[\gamma_1(k)] &= \Pr[\gamma_1^2(k)] = p_{11}, \\
\Pr[\gamma_0(k)] &= \Pr[\gamma_0^2(k)] = p_{00},
\end{align*}
$$

The probabilities $p_{01}, p_{11}, p_{21}, p_{00}, p_{10}$ and $p_{20}$ used in the simulation are summarized in Table 1.

The system matrix $\Phi$, which is equal to $A^T$, is calculated in (52). Also, from the state equation for $x(k)$ in (53), $K_x(k)$ is evaluated as

$$
K(0) = 0.202121, K(1) = 0.379147.
$$

Table 1 Probabilities for the Bernoulli variables $\gamma_0(k), \gamma_1(k), \gamma_2(k), \gamma_0(k), \gamma_0(k)$ and $\gamma_20(k)$.

<table>
<thead>
<tr>
<th>Cases of delay</th>
<th>Probability of the observation including both signal and observation noise</th>
<th>Probability of the observation not including signal and consisting of only observation noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>No delay</td>
<td>$\Pr[\gamma_0(k)] = 1 = p_{01}(k) = 0.8820$</td>
<td>$\Pr[\gamma_{00}(k)] = 1 = p_{00}(k) = 0.0180$</td>
</tr>
<tr>
<td>One-step delay</td>
<td>$\Pr[\gamma_1(k)] = 1 = p_{11}(k) = 0.0570$</td>
<td>$\Pr[\gamma_{10}(k)] = 1 = p_{10}(k) = 0.0030$</td>
</tr>
<tr>
<td>Two-steps delay</td>
<td>$\Pr[\gamma_2(k)] = 1 = p_{21}(k) = 0.0360$</td>
<td>$\Pr[\gamma_{20}(k)] = 1 = p_{20}(k) = 0.0040$</td>
</tr>
</tbody>
</table>

From (16) and (18), $\Phi^T$ and $K_x(0)$ are given by

$$
\Phi = \begin{bmatrix} 0 & 0.666667 & 0 & 0 & 0 & 0 \\ -0.8 & -0.266667 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},
$$

$$
K_x(0) = \begin{bmatrix} K_x(0) & \Phi K_x(0) & \Phi^2 K_x(0) \\ K_x(0) & K_x(0) & \Phi K_x(0) \\ K_x(0) & \Phi^2 K_x(0) & K_x(0) \end{bmatrix}.
$$

From (10) $\Phi_V$ and $K_V(L, L)$ are given by

$$
\Phi_V = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, K_V(L, L) = \begin{bmatrix} \bar{R} & 0 & 0 \\ 0 & \bar{R} & 0 \\ 0 & 0 & \bar{R} \end{bmatrix}.
$$
The fixed-point smoothing estimates $\hat{S}_1(k, k + \text{Lag})$ of $S_1(k)$, $\hat{S}_2(k, k + \text{Lag})$ of $S_2(k)$ and $\hat{S}_3(k, k + \text{Lag})$ of $S_3(k)$ are calculated recursively. Here, $\text{Lag}$ represents the fixed lag from $k + \text{Lag}$ to $k$. Fig.1 illustrates the fixed-point smoothing estimate $\hat{S}_1(k, k + 5)$ vs. $k$ for the white Gaussian observation noise $N(0, 0.1^2)$. Fig.2 illustrates the fixed-point smoothing estimate $\hat{S}_2(k, k + 5)$ vs. $k$ for the white Gaussian observation noise $N(0, 0.1^2)$. Fig.3 illustrates the fixed-point smoothing estimate $\hat{S}_3(k, k + 5)$ vs. $k$ for the white Gaussian observation noise $N(0, 0.1^2)$. Fig.4 illustrates the mean-square values (MSVs) of the filtering errors $S_1(k) - \hat{S}_1(k, k)$ and the fixed-point smoothing errors $S_1(k) - \hat{S}_1(k, k + \text{Lag})$ vs. $\text{Lag}$, $0 \leq \text{Lag} \leq 10$ , for the white Gaussian observation noises $N(0, 0.1^2)$, $N(0, 0.3^2)$ and $N(0, 0.5^2)$. For $\text{Lag} = 0$, the MSV of the filtering errors $S_1(k) - \hat{S}_1(k, k)$ is shown. In Fig.4, for each variance of the observation noise, the MSV of the fixed-point smoothing errors is larger than that of the filtering errors. Fig.5 illustrates the MSVs of the filtering errors $S_2(k) - \hat{S}_2(k, k)$ and the fixed-point smoothing errors $S_2(k) - \hat{S}_2(k, k + \text{Lag})$ vs. $\text{Lag}$, $0 \leq \text{Lag} \leq 10$, for the white Gaussian observation noises $N(0, 0.1^2)$, $N(0, 0.3^2)$ and $N(0, 0.5^2)$. Fig.6 illustrates the MSVs of the filtering errors $S_3(k) - \hat{S}_3(k, k)$ and the fixed-point smoothing errors $S_3(k) - \hat{S}_3(k, k + \text{Lag})$ vs. $\text{Lag}$, $0 \leq \text{Lag} \leq 10$, for the white Gaussian observation noises $N(0, 0.1^2)$, $N(0, 0.3^2)$ and $N(0, 0.5^2)$. In Fig.5 and Fig.6, there is a tendency that the MSVs of the fixed-point smoothing errors decrease as $\text{Lag}$ increases. Also, in Fig.5 and Fig.6, the estimation accuracy of the fixed-point smoothing estimate is superior to the filtering estimate. Here, the MSVs of the fixed-point smoothing and filtering errors are calculated by $\frac{\sum_{k=1}^{2000} (S_i(k) - \hat{S}_i(k, k + \text{Lag}))^2}{2000}, ~i=1,2,3$, respectively.
Fig. 1 Fixed-point smoothing estimate $\hat{S}_1(k, k+5)$ vs. $k$ for the white Gaussian observation noise $N(0, 0.1^2)$.

Fig. 2 Fixed-point smoothing estimate $\hat{S}_2(k, k+5)$ vs. $k$ for the white Gaussian observation noise $N(0, 0.1^2)$. 
Fig. 3 Fixed-point smoothing estimate \( \hat{S}_3(k, k+5) \) vs. \( k \) for the white Gaussian observation noise \( N(0, 0.1^2) \).

Fig. 4 Mean-square values of the filtering errors \( S_1(k) - \hat{S}_1(k, k) \) and the fixed-point smoothing errors \( S_1(k) - \hat{S}_1(k, k + \text{Lag}) \) vs. \( \text{Lag} \), \( 0 \leq \text{Lag} \leq 10 \), for the white Gaussian observation noises \( N(0, 0.1^2) \), \( N(0, 0.3^2) \) and \( N(0, 0.5^2) \).
Fig. 5 Mean-square values of the filtering errors $S_2(k) - \hat{S}_2(k, k)$ and the fixed-point smoothing errors $S_2(k) - \hat{S}_2(k, k + \text{Lag})$ vs. Lag, $0 \leq \text{Lag} \leq 10$, for the white Gaussian observation noises $N(0, 0.1^2)$, $N(0, 0.3^2)$ and $N(0, 0.5^2)$.

Fig. 6 Mean-square values of the filtering errors $S_3(k) - \hat{S}_3(k, k)$ and the fixed-point smoothing errors $S_3(k) - \hat{S}_3(k, k + \text{Lag})$ vs. Lag, $0 \leq \text{Lag} \leq 10$, for the white Gaussian observation noises $N(0, 0.1^2)$, $N(0, 0.3^2)$ and $N(0, 0.5^2)$.

5. Conclusions

In this paper, the RLS Wiener fixed-point smoother and filter are designed for the descriptor systems with randomly delayed, by multiple sampling intervals, or uncertain observations in linear discrete-time stochastic systems. In this paper, in addition to the multiply and randomly delayed observations [12], the uncertain observation in [13] is taken into account particularly for the linear discrete-time descriptor systems. The uncertain observation might correspond to the packet dropout. The packet dropout in the network systems is caused by the nodal delay as the sum of the processing delay, the queuing delay, the transmission delay and the propagation delay. Some numerical simulation results have shown that the devised estimators have feasible estimation characteristics.
Since the RLS Wiener estimators necessitate the information of the variance $Q$ of the input noise and the input matrix $\Lambda$ in the state equation (1), the estimation accuracy of the proposed RLS Wiener estimators are not degraded by the estimations of $Q$ and $\Lambda$.

References


1553 – 1559.


Housekeeping inspection and inventory analysis are the primary responses of engineering and logistics operations in hospitality industry. An intensive case study of professional research on Sheraton Gateway Hotel in Toronto Pearson International Airport, Toronto, Canada

*Greg MacNeil, Sheraton Gateway Hotel, Toronto, Canada

** Gazi Farok, York University and Sheraton Gateway Hotel, Toronto, Canada & Dhaka WASA, Bangladesh

Abstract

Housekeeping inspection maintains a chronological checklist and it has the major practice at the hospitality industry. Hospitality industry manages an imaging services to restaurants, lodging, event planning, theme parks, transportation, cruise line, tourism industry etc. Now days, hospitality as well as housekeeping are a dynamic and vibrant industry. It has different facilities of maintenances and direct operations with technology, engineering, housekeeping, kitchen workers, marketing, human resources, bartenders and supply chain logistic management.

Key words: Housekeeping, Engineering, Technology, Inspection, Operation.

1.0 Introduction

Sheraton was established in 1937 by Ernest Henderson and Robert Moore at Springfield, Massachusetts, USA. It is the one brand out of the thirty of Marriot + Starwood hotels and resorts from September 2016. It was the largest brand of Star Wood Hotels and Resorts. Right now, Sheraton has 435 hotels, 86 resorts in 70 countries. In Canada, Sheraton is operating 18 among the 84 Star Wood Hotels [Wikipedia, 2016]. Housekeeping performance (GEI) and engineering maintenance have established a good relationship and global reputation for Sheraton hotels.

1.1 Company Profile

The important business component for Sheraton Gateway, Toronto is hospitality by hotels and resorts. Sheraton Hotels and Resorts is a chain of luxury hotels owned by Marriott International. Sheraton has been welcoming guests through its doors since 1937 and it has grown to become one of the most popular, recognized and relied upon hotel brands in the world. It has more than 400 hotels and resorts located in some 70 countries around the globe. Sheraton Gateway is one of the most vibrant property which is linked to Toronto’s Pearson International Airport by a walk way, this modern hotel is 6.3 km from Centennial Park and 7.3 km from Woodbine Racetrack. It has relaxed rooms have WiFi (free), flat-screen TVs and desks, as well as mini-fridges and coffeemakers. Club rooms provide access to a lounge with complimentary continental breakfast, all-day snacks and afternoon appetizers, plus free WiFi. Its Amenities include an international grill, a bar and a lobby lounge, as well as a 24/7 fitness centre, a spa and an indoor pool. There are also 25 meeting rooms and a business center.

1.2 Services

General: Room Service, Restaurant, Pet Friendly, Air Conditioned, Mini Bar, Refrigerator, Cable / Satellite TV, Bath, Coffee / Tea Maker, Hair Dryer, TV, Shower, En suite, Private Bathroom, Spa Bath.

Activities: Fitness Room/Gym, Swimming pool, Massage / Beauty Centre, Sauna, Salon, Massage, Spa & Wellness Centre, Body Treatments, Pool Indoor, Business Center, Concierge, Elevator / Lift, 24-Hour Reception, Conference Room(s), Currency Exchange, Multilingual Staff, Porters, Wake-up Service, Meeting Rooms, Laundry service, Photocopier, Luggage Storage, Express Check-In/Check-Out.
Parking: There is an airport shuttle that runs from the hotel. Check-in: From 3:00 PM & Check-out: Prior to 12:00 PM.

At majority of Sheraton brand hotels, guests can earn a $5 credit at any of the hotel’s restaurants, or 500 Star points, for every night they opt out of housekeeping—for up to three consecutive nights (WSJ, 2010).

1.3 Objectives of the Study

It has been studied Sheraton Gateway websites, links and online documents and observed organizational chart and responsibilities of divisions and positions. It has performed a complete research and drawn an operational system which is coordinated by Director Operations at Sheraton Gateway. This research evaluated efficiency by analyzing the data and numerical calculations. It has gathered, analyzed, reported and summarized GEI (Guest Experience Index) scores to the Director of Operation. This study has implemented a supply chain, engineering reliability and logistics study with limitations and opportunities on operation and maintenance divisions. This research proposed a new operation strategy to obtain a balanced and sustainable economic system and assisted other departments for example engineering inspections, housekeeping with room inspections and tracking/maintaining inventories for lost and found, human resources with posters and many other various department specific duties. This study has organized and maintained files analyzed reports and surveyed with various trend analysis, and other duties as assigned. It is measured cost of logistics, SPG, Economic inventory model and connectivity of operation divisions. This study has calculated costs for various departments and quantifying where investments need to be made and if they can be made within the hotel and obtained a win-win participated management for purchasing with resource allocation. Finally, this research carried out and discussed ins and outs of operation systems with quality and risk analysis and recommended an updated operation strategy.

2.0 Comfortable Public Access and Win-Win Business

Sheraton Gateway provides direct access to Pearson Airport and features an indoor pool. Those staying at the property have access to a hair salon, a sauna and a beauty Centre. It has a fitness Centre with a swimming pool. Beauty treatments and massage are available at the hotel’s day spa. Rooms at Sheraton Gateway Hotel Toronto Airport come with mini bars, spa baths and refrigerators. Each has tea and coffee making facilities, a flat-screen TV and a private bathroom with a shower. It has relaxed bar offers an extensive drinks menu, while the in-house restaurant serves international cuisine. Breakfast is available each morning and can be served in the comfort of the rooms. It is within walking distance of Cara Operations. The hotel’s multilingual staffs are available to make sure that every traveler has a comfortable stay. Sheraton hotels formed the core of what came to be the ITT (International Telephone & Telegraph) Sheraton Luxury group, later Starwood’s Luxury Collection (Zagorin, 1993).

2.1 Organizational Chart: Sheraton Gateway, Pearson International Airport, Toronto, Canada
2.2 Every day Operations:  a) Technology - It contains wireless internet access, minimum bandwidth of 33 Mbps per 150 rooms (233 kbps per room) available for a guest internet use. Property web site might not conduct ecommerce transactions, games of chance or other regulated promotion activities b) Reservations: 7/24 hours for Gold, Platinum and SPG members; SPG check in station should be staffed first, c) Common Areas, d) Recreation, e) Guest Room, f) Food and Beverage, g) Meetings, h) Procurement, i) Welfare and Security, j) Engineering, k) Housekeeping, l) Banquet and Catering, m) Front Desk, o) Executive offices: Conference rooms, Meeting Arrangements.

Details operations:

i) Food Services: Catering, Banquet & Events, Restaurant bars

ii) Hotels: Revenue Management, Front Office, Rooms Division, Food & Beverages

iii) Events: Meeting & Conferences, Entertainment, Leisure, Sport Events

iv) Travel & Tourism: Cruise ships, Agencies & Tour Operations, Airlines

v) Luxury Services: Sales and Marketing, Customer Services, Luxury Brand Management

vi) Miscellaneous: Media, Finance and Real Estate, Education, Human Resources

Table 1: Routine operation criteria:

<table>
<thead>
<tr>
<th>No</th>
<th>4-Level Hotel Operation Services</th>
<th>No</th>
<th>4-Level Hotel Operation Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Technology</td>
<td>8.</td>
<td>Procurement</td>
</tr>
<tr>
<td>3.</td>
<td>Common areas</td>
<td>10.</td>
<td>Engineering</td>
</tr>
<tr>
<td>4.</td>
<td>Recreation</td>
<td>11.</td>
<td>Housekeeping</td>
</tr>
<tr>
<td>5.</td>
<td>Guest Room</td>
<td>12.</td>
<td>Banquet and Catering</td>
</tr>
<tr>
<td>6.</td>
<td>Food and Beverage</td>
<td>13.</td>
<td>Front Desk</td>
</tr>
<tr>
<td>7.</td>
<td>Meetings</td>
<td>14.</td>
<td>Executive Office</td>
</tr>
</tbody>
</table>

3.0 House Keeping Services: ABC OF Housekeeping (Audio/ Video knowledge):

A: Away with the old: Dirty glasses and bed room clean tools and chemicals by Garbage Bag

B: Bed: Corner fixing, top finishing, Make or straighten the bed as per brand.

* Bed-bugs training

C: Cleaning: Right chemicals at least 10 minutes to walk spray the bathroom clinic

D: Dusting: Cleaning chemicals and microfiber cloth.

E: Everything in the bathroom: Spray earlier; shining and polishing clearly

F: Finish the Bed Room

✓ Windows, doors and walls are cleaned
✓ Iron- no water
✓ Check the hangers
✓ Vacuum the whole room and all edges, corners should touch vacuum

Pls remember that it’s up to room attendant and supervisors to make the different. The shorter is the time of cleaning and tidying, the better is the efficiency, the hotel’s operation cost will decrease, and the hotel's competitiveness will ascend (Chen, 2001).

3.1 Inspections for Guest Room, Club Lounge etc (should be in excellent condition): Bed And Bedding, Headboard, Ceiling, Chair And Ottoman/Sofa, Closet Area, Iron, Board, Safe, Luggage Rack, Coffee Brewer, Ice
Bucket, Refreshment Center, Desk And Desk Chair, Entrance And Connecting, Drapes And Sheers, Dressers/Armoire, Floor, HVAC/ Vents, Lamps/ Lighting, Mirrors/ Artwork, Nightstands And Tables, Activity And Dining Tables, Coffee And End Tables, Telephone/ Clock Radio/ Television (s) and Remote, Walls and windows, Waste can/ Miscellaneous/ no offending door.

All equipment power and any related audio / video cables are neat and tidy and empty cable management techniques. In-Room Dining trolley and/or tray are clean on all surfaces and in all surfaces and all areas that are visible to the guest.


➢ No garbage in the room
➢ No hair in the bed
➢ No mold on the ties
➢ No hair in the bathroom
➢ Shuttle and other guest transport vehicles should be cleaned.

a) Inspection for Club Lounge, b) Inspection for Business center, c) Inspection for Lobby, d) Inspection for Corridors, Landing and Foyers, e) Inspection for Elevator, f) Inspection for Star wells, g) Inspection for Public restroom, h) Inspection for Fitness room, i) Inspection for Locker room/ Restroom, j) Inspection for Swimming Pool, h) Inspection for F & B Venue, i) Inspection for Ballroom, j) Inspection for Meeting Room, k) Inspection for Retail Shop, l) Inspection for Spa, m) Inspection for Outdoor common areas, n) Inspection for Pre-Function Areas, o) Inspection for all (office) heart of house areas, p) Inspection for all associate Lounges (heart of house areas), q) Inspection for all kitchen (heart of house areas), r) Inspection for all laundry (heart of house areas), s) Inspection for all loading dock (heart of house areas), t) Inspection for all locker (heart of house areas), u) Inspection for all rest rooms (heart of house areas), v) Inspection for all store and plant rooms (heart of house areas), x) Inspection for all entrances (heart of house areas).


4.0 Billing Systems

i) Flow diagram 1: Print function daily report

ii)
iii) Flow diagram 2: If payment is zero:

Event Order → Event Number → Open Check → Close check → Close

iv) Flow diagram 3: Payment is not zero ($54900)

Event Order → Order Number → Open Check → Post Check

    Print Report
    (A PDF document will open with print options) → Submit → Report

v) Flow diagram 4: Accounts for Consumption

Consumptions → No of water, No of Juice, No of Pops, No of Perrier → Open Check

    Print Report → Submit → Report → Post the check
vi) Flow Diagram 5: Operational Billing Master Accounts:

Engineering Maintenance & Logistics: Housekeeping department takes the first steps in maintenance functions for which the maintenance is ultimately responsible.

Manpower in Engineering Division: Manager (1)-Assistant Manager (1)-Coordinator (1)-Staff/ Mechanics (12) - 9 fulltime (6 mechanics + 3 painters), 2 part time mechanics, 1 HVAC technician

5.1 Inspecting of Synergy Voice: CMMS (Computer Maintenance and Management System)

➢ It is the Centre of Controlling System

How identify the complaints: A) Housekeeping Request, B) Maintenance Defects, C) Manager IT, D) Front Desk, E) Manager on Duty (MOD) or Guest Service Manager (GSM)

Complaints can be identified by:

i) Synergy (Employee ID) by Housekeeping and through the guest phone

ii) Star Guest (Switchboard) by Switch board agent/ Manager

5.2 Engineering Equipment:

i) Two cooling Tower (Each capacity 4000 Gallons)- Operated by Ontario Cooling Tower Services (OCTS) only for Summer ii) Glycol Plants (2 plants-getting cool by Glycol chemical)-3 pumps, 4 heating boilers, 2 coupling systems and 1 loop for whole building, ii) Mixing Chamber for tempered water (144 degree F), Capacity 50 Gallon Per Minute (GPM), iii) A Pumping system for guest rooms, meeting rooms and Banquet Halls, iv) Water Treatment Plants- 3 Plants, ALGACIDE-Cooling Tower Treatment (Liquid Microbicide) (2plants) and ECOPERSE (Cooling Tower Biological film cleaner) (1 plant) 2 boilers- approximately 75% heating capacity (Expansion Tanks)

➢ 1 boiler-magnetized- 95% capacity (Heat Exchanger)
Pre-season servicing (twice per year)

iv) Generator: Capacity (625 KVA/500KW), Power Factor-0.8, Volts-600/347, Phase-3, Hz-50, Duty-Standby, Test Run Quarterly (3 months), Manufactured- July, 1990,

Servicing by Toromont, Manufactured by: Cummins Ontario Inc,

6. Transformers (Switchgears Unit of Hotel): Capacity-27000 KVA/600KVA, Bus Amp-600 Amps, Cycles-60 C, Poles-3, Probability- it was made at beginning of the hotel

7. Reviewing and reshuffling of Sheraton Athletic Centre: Engineering design, arrangements of men and women practicing athletic areas and coordination of events.

5.3 Selected Projects: i) Sheraton Gateway Hotel, 8th floor club level lounge, Plan: Electrical

Terminal 3 Projects: Office Elevator Sections

Project Terminal 3: Parking Garage Foundation Plan

Project Terminal 3: Food Service Equipment Plan and Schedule.

Mechanical Plan- 8th floor club lounge, Sheraton Gateway Hotel-MOI


i) Part plans Mechanical pent house level: Observed correlation with structure, mechanical and electrical drawings,

ii) Electrical legend and single line diagram- Sheraton T3 Gateway Hotel and Fitness centre

iii) Sky lobby level electrical lighting & P.A. Speaker layout

iv) Sky lobby level electrical fire alarm layout,

v) Hotel Fire Protection System- Terminal 3 Project

vi) Fire Alarm System Retrofit 2008- Sheraton Gateway Hotel

6.0 Eye Viewer Project

Scope: i) Unobstructed door viewer on guestroom door (no scratches), ii) Unobstructed door viewer on guestroom door- must have a minimum of 160° (Unobstructed Angle- UA).
Table 2: Performance of Eye Viewers

<table>
<thead>
<tr>
<th>Room Number</th>
<th>Unit Status</th>
<th>Functionality</th>
<th>Phases</th>
</tr>
</thead>
<tbody>
<tr>
<td>324</td>
<td>New</td>
<td>Clear, $\text{UA} \geq 160^\circ$</td>
<td>OK</td>
</tr>
<tr>
<td>330</td>
<td>New</td>
<td>Clear, $\text{UA} \geq 160^\circ$</td>
<td>OK</td>
</tr>
<tr>
<td>334</td>
<td>Old</td>
<td>Cloudy and Scratch, $\text{UA} \leq 90^\circ$</td>
<td>Phase-1</td>
</tr>
<tr>
<td>360</td>
<td>Old</td>
<td>Cloudy and Scratch, $\text{UA} \leq 150^\circ$</td>
<td>Phase-2</td>
</tr>
<tr>
<td>478</td>
<td>Old</td>
<td>Cloudy and Scratch, $\text{UA} \leq 100^\circ$</td>
<td>Phase-1</td>
</tr>
<tr>
<td>523</td>
<td>Old</td>
<td>Cloudy and Scratch, $\text{UA} \leq 125^\circ$</td>
<td>Phase-2</td>
</tr>
<tr>
<td>507 (Accessible)</td>
<td>Old+ New</td>
<td>New is clear $\text{UA} \geq 160^\circ$ and old is Cloudy and Scratch $\text{UA} \leq 100^\circ$</td>
<td>New is OK and Phase-2 for old</td>
</tr>
</tbody>
</table>

WBS of Project: Total Room=474; 230 rooms newly installed and others will be replaced by phase-1 and phase-2.

Length of Project: Phase-1- Immediate and Phase-2- after finishing phase-1.

Sponsor/ Funding Organization: Marriot + Starwood Hotels and Resorts.

6.1 Data Analysis and Decision Making: Housekeeping Productivity, $G = \frac{(B-C-D-E-F)}{A} = \text{Total useful working time/ Time to clean for one guest room.}$

Table 3: Calculations for Housekeeping Productivity:

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>480</td>
<td>20</td>
<td>30</td>
<td>20</td>
<td>20</td>
<td>19.5</td>
</tr>
<tr>
<td>21</td>
<td>480</td>
<td>15</td>
<td>20</td>
<td>15</td>
<td>15</td>
<td>19.7619</td>
</tr>
<tr>
<td>25</td>
<td>480</td>
<td>18</td>
<td>22</td>
<td>18</td>
<td>16</td>
<td>16.24</td>
</tr>
<tr>
<td>22</td>
<td>480</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>19.09091</td>
</tr>
<tr>
<td>24</td>
<td>480</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>16.66667</td>
</tr>
</tbody>
</table>

18.3
A = Time to clean for one guest room (min)

B = Total Shift time (min)

C = Beginning of shift duties (min)

D = Morning Break (min)

G = Productivity Standard (Guest Room) in each shift

E = Afternoon Break (min)

F = End of shift duties (min)

Table 4: Performance Evaluation Criteria

<table>
<thead>
<tr>
<th>GEI (Guest Experience Index)</th>
<th>Evaluation Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factors</td>
<td>Month to date</td>
</tr>
<tr>
<td>Cleanliness</td>
<td>97.5%</td>
</tr>
<tr>
<td>Loyalty</td>
<td>93%</td>
</tr>
<tr>
<td>Housekeeping Productivity (Rooms per shift)</td>
<td>18.3</td>
</tr>
</tbody>
</table>

Figure 2: Relationships between Housekeeping and Engineering Complaints

Complaints (272): Housekeeping*(Bathroom, Bedroom, Garbage and Smell) & Engineering** (Noise and Maintenance)

- Bathroom* (6.94%)
- Bedroom* (13.89%)
- Garbage* (9.72%
- Smell* (28.8%)
- Noise** (8.33%)
- Maintenance** (32.41%)

6.2 Project Review: Lobby Entrance: Title: HVAC and Sprinkler layout specifications- Mechanical Installed: 2000/02/24, Reviewed: 2016/10/06

Descriptions: i) Reviewed fabrication of all ductwork and hangers to ASHRAE & SMANCA recommendations, ii) Ran all ductwork and piping as high as possible, iii) Located thermostat where shown on plan, iv) Insulated all concealed Supply Air ductwork with 1-inch Fibre glass insulation complete with foil faced vapour barrier, v) Reviewed and supply and install all necessary balancing and volume dampers. Balanced all air systems to air volume noted on the drawings.
Table: 5 Calculations of cooling load and heating load: a) Cooling Load Calculations

<table>
<thead>
<tr>
<th>Components</th>
<th>Weightage</th>
<th>Total (Btuh)</th>
</tr>
</thead>
<tbody>
<tr>
<td>South Glass</td>
<td>240x165x0.53x0.8x0.5</td>
<td>8,395</td>
</tr>
<tr>
<td>East Glass</td>
<td>280x214x0.29x0.8x0.5</td>
<td>6,951</td>
</tr>
<tr>
<td>West Glass</td>
<td>280x214x0.04x0.8x0.5</td>
<td>9,587</td>
</tr>
<tr>
<td>Glass</td>
<td>800x0.62x17</td>
<td>8,432</td>
</tr>
<tr>
<td>Wall</td>
<td>350x0.1x17</td>
<td>595</td>
</tr>
<tr>
<td>Lights</td>
<td>1.00wx3413</td>
<td>3,413</td>
</tr>
<tr>
<td>Peoples</td>
<td>7x250</td>
<td>1,750</td>
</tr>
<tr>
<td>Peoples</td>
<td>7x250</td>
<td>1,750</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>40,873 Btuh</td>
</tr>
</tbody>
</table>

Air Temperature = 40,873/(1.08x2000 cft) = 18.9°F

b) Heating Load Calculations:

<table>
<thead>
<tr>
<th>Components</th>
<th>Weightage</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glass</td>
<td>800x0.62x70</td>
<td>34,720</td>
</tr>
<tr>
<td>Walls</td>
<td>350x0.1x70</td>
<td>2,450</td>
</tr>
<tr>
<td>Floor</td>
<td>500x0.5x70</td>
<td>17,500</td>
</tr>
<tr>
<td>Infiltration</td>
<td>40x15cfmx1.08x70</td>
<td>45,360</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>100,030Btuh</td>
</tr>
</tbody>
</table>

Heating (Furnace) Operation Temperature = 100,030/(1.08x2000 cft) = 46.3°F

6.3 Sheraton Hotel Emergency Supplies Inventory

Table 6: List of Emergency Supplies Inventory at Sheraton Gateway, Pearson Airport, Toronto, Canada:

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>Name of the Item</th>
<th>Number of Items</th>
<th>Specification of Items</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Bolt Cutter</td>
<td>04</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Quantity</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>--------------------------------------------------</td>
<td>----------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Duracell Alkaline Batteries-D</td>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Dehumidifier Console</td>
<td>01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Preventive Maintenance – Elevator</td>
<td>02</td>
<td>Thysseukrupp Elevator</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Canners and Handlers</td>
<td>01 packet</td>
<td>Product of Thailand</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Flash Lights (Mglite)</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>N95 Particulate Respirator (Mask)</td>
<td>4 Cartons</td>
<td>N95</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Ice Cleaner (plough)</td>
<td>02</td>
<td>Garant</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>RIDGID (dual flek locking hose)/Performance Filter</td>
<td>03 Cartons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Protective Footwear</td>
<td>06 Pairs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Cables (Yellow)</td>
<td>02 coils</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Cables (Red)</td>
<td>06 coils</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Cables (Green)</td>
<td>0.5 coil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Scotchilite-FRR 1500 Lime-L</td>
<td>04 Pcs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Venues Emergency Supplies</td>
<td>09 Bags</td>
<td>Each bag: Small flash lights 10 nos, Large flash light 01 no, Blood Borne clean up 01 no, Portable radio 01 no, Safety vest 01, Rubber Gloves 01 no, Masks 10 no, Goggles 01 no, Whistle 01 no, First and Kit 01 no, Ice Pack 01 no.</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Rechargeable clock/ radio/light</td>
<td>03 Cartons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Flash light stickers (6 inches)</td>
<td>25 Packets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Ice remover (sable)/ Garant</td>
<td>04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Filter (Sucking / Exhaust)</td>
<td>02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Temporary stand for posters, notices, flyers</td>
<td>02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Decoration materials ( with electrification)</td>
<td>04 bags</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table-7: List of Sheraton Brand Standard Emergency Supplies

<table>
<thead>
<tr>
<th>Sl. No</th>
<th>Item</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Batteries-9 volt</td>
<td>4</td>
</tr>
<tr>
<td>2.</td>
<td>Batteries-AAA</td>
<td>12</td>
</tr>
<tr>
<td>3.</td>
<td>Batteries-AA</td>
<td>12</td>
</tr>
<tr>
<td>4.</td>
<td>Batteries-C</td>
<td>10</td>
</tr>
<tr>
<td>5.</td>
<td>Batteries-D</td>
<td>10</td>
</tr>
<tr>
<td>6.</td>
<td>Batteries-rechargeable for 2 way radios</td>
<td>1 extra per radio</td>
</tr>
<tr>
<td>7.</td>
<td>Blankets/ pillows/ cots</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Bolt Cutter</td>
<td>1</td>
</tr>
<tr>
<td>9.</td>
<td>Built-horn</td>
<td>1</td>
</tr>
<tr>
<td>10.</td>
<td>Duct and masking tape</td>
<td>2 rolls each</td>
</tr>
<tr>
<td>11.</td>
<td>Emergency vests</td>
<td>5</td>
</tr>
<tr>
<td>12.</td>
<td>Fire extinguishers-extras</td>
<td>3</td>
</tr>
<tr>
<td>13.</td>
<td>First and Kit</td>
<td>2</td>
</tr>
<tr>
<td>14.</td>
<td>Flashlights</td>
<td>2</td>
</tr>
<tr>
<td>15.</td>
<td>Fluorescent tape or rope</td>
<td>1 roll tape/ 40’ rope</td>
</tr>
<tr>
<td>16.</td>
<td>Foul weather Gear</td>
<td>2 sets</td>
</tr>
<tr>
<td>17.</td>
<td>Light sticks</td>
<td>2 dozen</td>
</tr>
<tr>
<td>18.</td>
<td>Master keys</td>
<td>1 extras</td>
</tr>
<tr>
<td>19.</td>
<td>Plastic bags, large</td>
<td>1 box</td>
</tr>
<tr>
<td>20.</td>
<td>Radio- AM/ FM radio- battery operated</td>
<td>1</td>
</tr>
</tbody>
</table>

7.1 Conclusions

This intensive research is carried out to know hotel industries and its various departments, activities and their relationships with functioning of a good working system. Usually Sheraton annually implements a new practice to improve financial viability and guest and employee satisfaction (Judy & Cathy, 1999). This field study has created to learn specific opportunities for obtaining a specialization on supply chain management, logistics and operations at Marriot + Starwood Hotel and resorts.

7.2 Recommendations:

An international standard hotel is used to evaluate value critically linked to the comfort and security of its all level guests. From the lobby to the presidential suite, from the parking garage to room access, universal solutions are focused and recommended the following issues on what matters most to the hospitality industry [www.climatec.com]:

279
i) Inventory management, engineering maintenance and housekeeping inspections have common complain objects and performances, i.e.; economic order quantity for food and beverages, operations of electrical, mechanical, plumbing, carpentry, civil and boiler works etc.

ii) Long inspection form for room inspection has better application to identify minor error for improving housekeeping cleanness ratio as well as GEI etc

iii) This resourceful study will be an asset to obtain a job description of management trainee position at Marriot + Starwood Hotel and Resorts as well as ins and outs understanding of hospitality industry.

8.0 References


4. Judy A. Siguaw & Cathy A. Enz, (1999), Best Practices in Hotel Operations, Cornell University School of Hotel Administration, USA


Appendix

Standardized Organizations

✓ ASTM- American Society for Testing Materials
✓ ASA- American Standard Association
✓ AKI- Architectural Woodwork Institute
✓ CEMA- Canadian Electrical Manufactures Association
✓ CGSB- Canadian General Standards Board
✓ CSA- Canadian Standards Association
✓ NBC- National Building Code of Canada
Make Semantic Analysis of Opinions about social networking using Blog Search Engines

Abstract

There is no consensus amongst the academic communities whether Social media is a boon or bane for the students. Semantic analysis and blog search was used to get extracts opinions of the bloggers regarding various aspects of using social media in education and employment.

Indexing terms/Keywords: social media, blog search engine, semantic analysis.

Introduction

The Semantic Analysis can be understood as the process of finding language-independent meaning from written contents. Blogs are online platforms for individuals to express their opinion about events, issues or person etc and publish it on world wide web. Blog search Engines is a special purpose search Engines used to search blogs. Use and abuse of social networking sites is in news day in and day out. There is no consensus amongst the academic communities whether this technology is a boon or bane for the students. In order to find out the authors conducted a survey amongst teachers and students of higher education and the results was published in [1]. However it was realized that survey based methods have their own limitation. In order to reduce the bias, it was decided to use semantic analysis approach to extract opinions from blogs.

Materials and Methods

Different groups of Keywords were related to the research topic were chosen like

- social networking sites
- Problems of social networking sites
- Prospects of social networking sites
- use of social networking sites in education
- use of social networking sites in employment

The keywords their synonyms and variation were formed For e.g. social networking, social network, social networking site can be different variation of social networking sites. Demerits, disadvantage, limitations, drawbacks were different synonyms of problems. These variation and synonyms when used in different permutation and combination resulted into 202 keywords. These 202 keywords were searched in 3 different blog search engines i.e Google blog search, Technorati, & Icerocked and hit statistic were noted. This search resulted into 606 rows of data. This data was organized as follows.

<table>
<thead>
<tr>
<th>Sr.No</th>
<th>Keyword</th>
<th>Blog search engine</th>
<th>Output</th>
</tr>
</thead>
</table>

Results and Discussion
### Table 2: Finding of the Research

<table>
<thead>
<tr>
<th>Keywords</th>
<th>No of Variations/Synonyms</th>
<th>Google Blogs search</th>
<th>Ice rocket</th>
<th>Technorati</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prospects</td>
<td>78</td>
<td>1416736200</td>
<td>289039</td>
<td>1078</td>
</tr>
<tr>
<td>Problems</td>
<td>47</td>
<td>1103221200</td>
<td>154733</td>
<td>681</td>
</tr>
<tr>
<td>Education</td>
<td>28</td>
<td>4412603600</td>
<td>15961</td>
<td>38</td>
</tr>
<tr>
<td>Employments</td>
<td>49</td>
<td>1190034000</td>
<td>12412</td>
<td>17</td>
</tr>
<tr>
<td>Grand Total</td>
<td>202</td>
<td>8122595000</td>
<td>472145</td>
<td>1814</td>
</tr>
</tbody>
</table>

A two factor ANOVA without replication was conducted and responses from different blog search engines taking keywords as dimension. This methodology was adopted from Khosla and Acharya (2011) where the author used different search engines and medical streams as dimensions [3]. The output is reported in Table 4 and Table 3:

### Table 3: Analysis of the Variance

**ANOVA**

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUMMARY</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prospects</td>
<td>4</td>
<td>1.42E+09</td>
<td>354256598.8</td>
<td>5.02E+17</td>
</tr>
<tr>
<td>Problems</td>
<td>4</td>
<td>1.1E+09</td>
<td>275844165.3</td>
<td>3.04E+17</td>
</tr>
<tr>
<td>Education</td>
<td>4</td>
<td>4.41E+09</td>
<td>1103154907</td>
<td>4.87E+18</td>
</tr>
<tr>
<td>Employments</td>
<td>4</td>
<td>1.19E+09</td>
<td>297511619.5</td>
<td>3.54E+17</td>
</tr>
<tr>
<td>No of Variations/Synonyms</td>
<td>4</td>
<td>202</td>
<td>50.5</td>
<td>425.6667</td>
</tr>
<tr>
<td>Google Blogs search</td>
<td>4</td>
<td>8.12E+09</td>
<td>2030648750</td>
<td>2.54E+18</td>
</tr>
<tr>
<td>Ice rocket</td>
<td>4</td>
<td>472145</td>
<td>118036.25</td>
<td>1.74E+10</td>
</tr>
<tr>
<td>Technorati</td>
<td>4</td>
<td>1814</td>
<td>453.5</td>
<td>268309.7</td>
</tr>
</tbody>
</table>

### Table 4: ANOVA Results
### Main Text (Review only)

This section may be divided into subsections or may be combined.

### Conclusions

This is evident from table 3 and table 4 the $F$-value for row as well as $F$ value for column is greater $P$-value and $F$ Critical. However, it can see that $F$-Value is column is greater than $F$ value for columns. It can say that more people blogging about prospects as compared to problems.

### Data Availability (excluding Review articles)

The data can be obtained from the authors on request. The excel sheet or link to the google document will be sent.

### Conflicts of Interest

Not applicable

### References


Designing a Customer Relationship Management System in Online Business

Negin Ziaei\textsuperscript{1}, Hamed Fazlollahtabar\textsuperscript{2}

\textsuperscript{1}Department of Information Technology Engineering, Mazandaran University of Science and Technology, Babol, Iran

\textsuperscript{2}Department of Industrial Engineering, School of Engineering, Damghan University, Damghan, Iran

hfazl@du.ac.ir

Abstract

With the advancement of online shopping technology, it has become the first choice for most consumers. The activity of online stores in this competitive business space should be in line with the expectations of their customers. Understanding, collecting, maintaining and organizing data in online stores makes it easier for managers to decide. So, in this research, we examine the textual and non-textual of user opinions and reviews. We use rapid miner software and text mining. In this research, the processes are aimed at finding active users, analyze the user type and their suggestions, analyzing the strengths and weaknesses of the products, and categorizing them with the K-NN and Naïve Bayes algorithms. Finally, suggestions were made to increase loyalty and improve business using the results obtained from the processes.

Keywords: Business intelligence (BI); Customer relationship management (CRM); Online shopping

Introduction

Interest in customer relationship management (CRM) began to grow in 1990s (Ling & Yen, 2001; Y. Xu, Yen, Lin, & Chou, 2002). Regardless of the size of an organization, businesses are still motivated to adopt CRM to create and manage the relationships with their customers more effectively. An enhanced relationship with one’s customers can ultimately lead to greater customer loyalty and retention and, also, profitability (Ngai, 2005). CRM has been one of the greatest technological contributions to enterprises in the twenty-first century (Chao, Jen, Chi, & Lin, 2007). CRM consists of guidelines, procedures, processes and strategies which provide organizations the ability to merge customer interactions and also keep track of all customer-related information. Technologies are utilized to attract new and profitable customers, retain and strengthen ties with current ones. CRM revolves around the concept of maintaining long-lasting, valuable relationships with customers (Khan, Ehsan, Mirza, & Sarwar, 2012).

Online shopping is growing faster than any moment in history of commerce and attracting many consumers to choose shopping online instead of traditional shopping channels. This new type of shopping mode, also known as e-shopping, online shopping, network shopping, Internet shopping, or Web-based shopping, featuring in freeing consumers from having to personally visit physical stores, is anticipated to greatly change people’s everyday lives (Hsiao, 2009). The importance of effective CRM implementation is intensified in the e-business environment since customer loyalty is much more difficult to establish in this domain (Kimiloğlu & Zaralı, 2009).

The simple definition of e-CRM is customer relationship management on the web; however, e-CRM also includes the use of e-mail, e-commerce activity, and any other Internet-based customer touch points. Electronic customer relationship management (e-CRM) enables retailers to better meet the needs of their customers across retail formats and, at the same time, to maximize the strategic benefits of a multichannel strategy (Warrington, Hagen, & Feinberg, 2009).

Till now, the concepts of satisfaction and loyalty for website which involved in providing services on the website and transacting online, is a central concern of marketers. In recent years, electronic commerce has entered a phase of exponential growth and the use of the Internet in the consumer decision making process ensures that traders to make greater use of this tool. While consumer behavior in e-commerce seems to be a complex subject,
the consumer expectations are changing, challenging traditional patterns of supply of commercial websites (Bashar & Wasiq, 2013).

In this research, we examine the textual and non-textual of user opinions and reviews about products they buy. We use rapid miner software and text mining and Data mining techniques to turn these data into understandable knowledge. The processes are aimed at finding active users, analyze the user type and their suggestions, analyzing the strengths and weaknesses of the products, and categorizing them. Finally, suggestions were made to increase loyalty and improve business using the results obtained from the processes.

The rest of this paper is organized as follows. Next, the related literature is reviewed. In Section 3, the proposed problem is described. Section 4, illustrates the algorithms required for implementation. Section 5 conducts a case study. Finally, we conclude in Section 6.

Related works

Chen and Tseng (2011) propose a method for evaluating the quality of information in product reviews. They treat the quality evaluation of product reviews as a classification problem and employ a multiclass support vector machine (multiclass SVM) model to categorize reviews. In addition, they adopt a mature information quality (IQ) framework. Ghose and Ipeirotis (2011) study, integrates econometric, text mining, and predictive modeling techniques. By using Random Forest-based classifiers toward a more complete analysis of the information captured by user-generated online reviews in order to estimate their helpfulness and economic impact. Many tourism companies now actively use Internet sites as a key marketing and sales vehicle for their products and services. To be successful, tourism e-commerce services must be trustworthy. Zheng et al. (2013) developed a semi-supervised system called Online Review Quality Mining (ORQM). Embedded with independent component analysis and semi-supervised ensemble learning, ORQM exploits two opportunities: the improvement of classification performance through the use of a few labeled instances and numerous unlabeled instances, and the effectiveness of the social characteristics of e-commerce communities as identifiers of influential reviewers who write high-quality reviews.

Faed et al. (2014) proposed a conceptual framework including mathematical models, hypothesised relationships, perceived value and interactivity between customer, business and the system, as well as customer satisfaction and loyalty analytics. Based on nonlinear modelling and using a fuzzy inference system, namely the Takagi–Sugeno-type approach, they defined fuzzy rules, by means of which they ascertain the relationship between customer satisfaction and the main relevant variables. Jack and Tsai (2015) presented a framework for using text mining and R, a statistical software to gather customer feedback from Amazon website. A case study comparison of three devices compares and contrasts positive and negative aspects mentioned by the users, which is useful to improve future generations of products. Dixit and Kr (2016) analyzed feedback of customer on three different mobile brand for this two different classifiers has built to extract the feedback of customers, those are shared in e-commerce website and classify them broadly into 3 categories – good, bad and mixed. Xu and Li (2016) used latent semantic analysis (LSA), which is a text mining approach, they analyze online customer reviews of hotels. Their study provides a clue for hoteliers to enhance customer satisfaction and alleviate customer dissatisfaction by improving service and satisfying the customers’ needs for the different types of hotels the hoteliers own. Małeckiet al. (2017) proposed the methodological approach of the decision support system for identifying Internet customer typology. Online users clustering was performed by cluster analysis. Additionally, by introducing a prediction mechanism based on the mathematical model of the Graph Cellular Automaton, the new customer can be quickly adjusted to the defined group of shop customers.

Proposed problem

With the advancement of technology, people are looking for the easiest way to do their daily routines that online shopping is one of them. There are many reasons to shop online instead of going to the stores. It’s easier to shop online, which means that it’s always available 24 hours a day, 7 days a week, and in any weather conditions without having to look for a parking space and suffering in traffic these days. One of the other reasons is ease of use for searching the Intended products and selecting between various brands and the ability to compare products and their prices for reaching a good decision to purchase. So, given the benefits of online shopping, it can be concluded that there is a lot of competition between online stores. In this business,
customers play an essential role. In fact, choosing products online is consumers and customers job. They have the power to choose and the world of online retailers at their fingertips. If online shops do not meet their needs, they quickly go to the rival store. This is where the issue of customer loyalty is raised, which is also the goal of our research to increase loyalty in online shopping.

Increasing customer loyalty plays an important role in the organization to be successful, especially when customer purchases are not long-term success. To measure customer loyalty in a reviewing mechanism, we analyze the opinions, reviews, and ratings of customers and consumers who have shared their experiences on the website. Comments and ratings of products on the website are an important source of information for customer decision making. Here, our data is mostly made up of comments and customer reviews that are unstructured texts. So, we use text mining techniques in this research and analyze consumer’s reviews by creating processes in rapid miner. The steps involved in this research to reach the goal and to analyze the review of the users are briefly illustrated as a model in Figure 1.

In the process of buying from online shops, users search on different parts of a website to find the product they wants. In this process, the user will be faced with a variety of options, and they may Giving up a lot of high quality products with small number of comments. In this research, finding the desired product for analysis is considered as an early and important step. At this stage, we have tried to use the best-selling, most viewed and most popular products that also have a large number of comment.

Figure 1 Research model
After selecting the desired products, we will extract and translate them for the next step. In this research, we use the web content mining method to extract data from comments. Web Scraping (also termed Screen Scraping, Web Data Extraction and Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format. We use data miner and google translate extension in Google Chrome, for web scraping and translate comment (Persian to English) at the same time.

**Required algorithms**

In data mining processes for prediction, we need to convert text data to vector and numeric data. This is done by the TF-IDF method. After the data are pre-processed then they are ready to use them in RapidMiner processes and data mining algorithms. Some processes include classifications predicting some features of products. In these processes, we use K-NN and Naive Bayes algorithms that described below.

**TF-IDF**

The technique of calculating text weighting is called TF-IDF, which stands for Term Frequency–Inverse Document Frequency.

Calculating TF is very easy: it is simply the ratio of the number of times a keyword appears in a given document, \( n_k \) (where \( k \) is the keyword), to the total number of terms in the document, \( n \):

\[
TF = \frac{n_k}{n}
\]

Considering the above example, a common English word such as “that” will have a fairly high TF score and a word such as “RapidMiner” will have a much lower TF score. IDF is defined as follows:

\[
IDF = \log_2 \frac{N}{N_k}
\]

Where \( N \) is the number of documents under consideration. For most text mining problems, \( N \) is the number of documents that we are trying to mine, and \( N_k \) is the number of documents that contain the keyword, \( k \). Again, a word such as “that” would arguably appear in every document and thus the ratio \( \frac{N}{N_k} \) would be close to 1, and the IDF score would be close to zero for “that.” However, a word like “RapidMiner” would possibly appear in a relatively fewer number of documents and so the ratio \( \frac{N}{N_k} \) would be much greater than 1. Thus the IDF score would be high for this less common keyword (Kotu & Deshpande, 2014).

Finally, TF-IDF is expressed as the simple product as shown below:

\[
TF-IDF = n_k / n \times \log_2 \frac{N}{N_k}
\]

In RapidMiner studio user has four different options, each of which represents the relationship between the words/terms and the documents with different numbers:

1. **Binary Term Occurrence** places 1 in the intersection cell between a document (row) and a word/term (column) if the word/term occurs at least once in that document and places 0 otherwise. The number of occurrences in the document is ignored in this measure.

2. **Term Occurrence** places the exact number of occurrences of a word/term in the intersection cell between the document (row) and the word/term (column). If the word/term does not occur in that document, 0 is placed in the intersection cell.

3. **Term Frequency** places the relative frequency of the word/term in the document in the intersection cell. This measure is calculated by dividing the number of occurrences of a word/term into the number of total words in that document.
4. **TF-IDF** stands for *Term Frequency-Inverse Document Frequency*. It is arguably the most commonly used numerical representation in text mining. It calculates a numerical value that emphasizes both the frequency of the term in a document (more is better) and the rareness of the same term in the collection of all documents (less is better) (Miner, Elder IV, & Hill, 2012).

In this study, we use **TF-IDF** and **Term Frequency** for text mining operations.

**Naïve Bayes**

The naïve Bayesian algorithm is built on Bayes' theorem, named after Reverend Thomas Bayes. We make use of Bayesian algorithm for the conditional stochastic text occurrence. Consider \( X \) as an evidence and \( Y \) as an output. Then, the probability of output \( P(Y) \) is called the *prior probability*. \( P(Y|X) \) is called the *conditional probability*, which provides the probability of an outcome given the evidence when we know the value of \( X \). Bayes' theorem states that

\[
P(Y|X) = \frac{p(Y) \cdot P(X|Y)}{P(X)}
\]

More generally, in an example set with \( n \) attributes \( X = \{X_1, X_2, X_3... X_n\} \),

\[
P(Y|X) = \frac{p(Y) \cdot \prod_{i=1}^{n} P(X_i|Y)}{P(X)}
\]

Since \( P(X) \) is constant for every value of \( Y \), it is enough to calculate the numerator of the equation \( p(Y) \cdot \prod_{i=1}^{n} P(X_i|Y) \) for every class value (Kotu & Deshpande, 2014).

4.3. **K-NN**

The \( k \)-nearest-neighbor method was first described in the early 1950s. The method is labor intensive when given large training sets and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition. When given an unknown tuple, a **\( k \)**-nearest-neighbor classifier searches the pattern space for the \( k \) training tuples that are closest to the unknown tuple. These \( k \) training tuples are the \( k \) “nearest neighbors” of the unknown tuple.

“Closeness” is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, \( X_1 = (x_{11}, x_{12}, ... , x_{1n}) \) and \( X_2 = (x_{21}, x_{22}, ... , x_{2n}) \), is

\[
dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}
\]

In other words, for each numeric attribute, we take the difference between the corresponding values of that attribute in tuple \( X_1 \) and in tuple \( X_2 \), square this difference, and accumulate it. The square root is taken of the total accumulated distance count. Typically, we normalize the values of each attribute before using above Equation. This helps prevent attributes with initially large ranges (e.g., *income*) from outweighing attributes with initially smaller ranges (e.g., binary attributes). Min-max normalization, for example, can be used to transform a value \( v \) of a numeric attribute \( A \) to \( v' \) 0 in the range \([0, 1]\) by computing

\[
v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A}
\]

where \( \text{min}_A \) and \( \text{max}_A \) are the minimum and maximum values of attribute \( A \). For \( k \)-nearest-neighbor classification, the unknown tuple is assigned the most common class among its \( k \)-nearest neighbors. When \( k = 1 \), the unknown
tuple is assigned the class of the training tuple that is closest to it in pattern space. Nearest-neighbor classifiers can also be used for numeric prediction, that is, to return a real-valued prediction for a given unknown tuple. In this case, the classifier returns the average value of the real-valued labels associated with the $k$-nearest neighbors of the unknown tuple (Han, Pei, & Kamber, 2011).

**Case study**

The purpose of this research is to analyze the texts of customers’ comments with the help of text mining and provide strategic Approach for online shops to enhance customer loyalty. Digikala is the online shop that we investigated for turn their comments into knowledge. The segmentation goods on the digikala website includes 8 parts: digital, fashion and clothing, cosmetics, books, culture and art, sports and travel, mother and child, and vehicles and industrial. According to the reviews, we are select products from the digital category. Some reviews examples include:

- In the fashion and clothing section, most of the products belonged to another site called Digi style. We noticed that users did not comment on Digi style website, even though there was a definite panel for the review section. In order to increase the trust and attraction of customers, to improve customer relationship management, the client should be encouraged to interact with the site. Also, by reviewing this section, we found that bestsellers, as well as the most popular products are smart watches and gadgets, could be considered as part of digital products.

- In the home and kitchen section, all the products have small amount of comment except audio and video equipment and game consoles.

- In the vehicle and industrial section, there were few products, which there have small amount of a comments and those a lot of comments, have nobuyers.

As discussed above, digital products have more comments than other section products so for this study we select 4 products from this section. Categories of the products we intend to mind their comments are: smartphones, tablets, external hard and speakers that shown in Table 1. Comments were sorted in the order of their usefulness, not according to the date.

After selecting the products, we are at the stage of extracting customer data and comments about the four products that mentioned before. With the web scraping technique and using the data miner extension, 4 recipes were created for selected products.

<table>
<thead>
<tr>
<th>Table 1 selected products</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>products</strong></td>
</tr>
<tr>
<td>Samsung Galaxy Note 8 SM-N950FD Dual SIM 64GB Mobile Phone</td>
</tr>
<tr>
<td>Product</td>
</tr>
<tr>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>ASUS Zen Pad 8.0 Z380KNL 4G 16GB Tablet</td>
</tr>
<tr>
<td>Western Digital Elements External Hard Drive - 2TB</td>
</tr>
<tr>
<td>JBL Go Portable Bluetooth Speaker</td>
</tr>
</tbody>
</table>

After initial preparation and cleaning, data are entered into the RapidMiner software or RM. RM consists of two important section view, design view which processes are formed and displays and result view that show the results of that processes.

In this research, five processes are constructed using RM that shown in Figure, which we will explain about them in further detail. Process 1 and 2 are just generate inputs for other process so we are not discuss about them.
Figure 2 Research process

Figure shows the results view and the chart section of a scatter diagram of output data from p3. The x-axis specifies the number of times each user comment, including 4 values, but because of the jitter data are in decimal form. The Y-axis shows the date. The colors in this chart are the authors. The comments in the shape of the circle are all green, which means the authors are identical. The number of times the author comment is in x-axis and the dates can also be distinguished from the y-axis.

Figure 3 Scatter chart of output data (P3)
The results of this process include 2 (Persian and English) example set. The output of p3 is table consists of 157 rows, from 1278 review 175 reviews were made by guest users and 67 of them were known as active user. Now, in Table 2, we discuss about these 67 people, which comments 157 times.

**Table 2** Analyze active user comments

<table>
<thead>
<tr>
<th>num</th>
<th>observation</th>
<th>analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The number of dispersion of buyer and non-purchasers, except mobile, which was only one buyers that comment twice, was regular in other products.</td>
<td>Despite the fact that the comments in the mobile category are also much more in terms of number, detail and accuracy of product reviews than other categories, this is a sign of uncertainty and trust of users in expensive items. So, there is a need to attract customer loyalty and trust in these types of products.</td>
</tr>
<tr>
<td>2</td>
<td>Some people commented on several different products.</td>
<td>The presence of people in Group [2] is a good sign that users are constantly sharing their opinions when they buy and after using the product at different period of times, which has a significant impact on users who are going to buy that products. Somehow it will attract and increase the range of new customers.</td>
</tr>
<tr>
<td></td>
<td>Many comments have been posted during a period of time for months or even years, including complete analysis and reviews.</td>
<td>In fact, these users as well as the users of Group [1] are active users who need more attention and more services with the aim of maintaining them and increasing their loyalty.</td>
</tr>
<tr>
<td></td>
<td>Some other comments were also written in one day, sometimes containing similar content or repetitive texts.</td>
<td>Users of the group [3] have little or no effect on business improvement, it is possible to increase this impact by giving them more awareness.</td>
</tr>
</tbody>
</table>

**3**

- Despite having a separate section called Q&A, some questions have been expressed in some comments!
- Some comments also include requests for the availability of goods, the existence of a particular color, and a special offers.

- Changing the UI/UX website to help guide the user towards the Q&A section or add the ability to reply to comments in the same section of the comments may help to solve the issue.
- By adding a part to express user requests, this weakness can turn out to be a strong point, which will greatly increase sales.

The output of the k-nn algorithm model in Figure shows that the learning model consists of 895 comments and 50133 columns with n = 8, and hard, mobile, speaker and tablet classes.
Figure 4 The model output of the \( k\)-nn algorithm (P4)

Example Set output of this algorithm contains 383 comments that have our test data, 50133 regular attributes and 8 special attributes. These eight attributes are, id column with the role id, a category column that represents the actual commented class, and another prediction (category) column that displays the predicted classes, we also have 4 confidence columns for each label class, which indicates the likelihood of each label class value.

The output of the model in the Naïve Bayes algorithm has 3 tabs, as shown in Figure 4. The first one is description shows that the output of the model has 23832 columns, and the numbers generated for each class represent the probability of data in that class. Given that the number of comments in the mobile class is greater than the other 3 classes, the probability is also higher. The next tab is chart, contains probability density functions for the specified attributes. In these charts, the x-axis is the value or in other words, the weight of the word, and the y-axis represents the density of that value in different classes with the color of that class. The third tab is distribution table with all attribute values and corresponding probability measures. In this table, the mean parameters and standard deviations of each word are calculated in different classes. For example, in distribution table the mobile column is set to ascending and the note word is in the highest row, so the number of occurrences is also high in this class. In the chart section, we also set attr to note, in the specified curves it is known that the part where the note is zero is the highest density for 3 classes: hard, speakers and tablet, and the green curve showing the mobile class has densities in non-zero sections indicates The existence of this word is in the mobile class and it is almost non-existent in the other 3 classes.
In Figure 5, the accuracy and classification errors of the comments are also expressed in terms of mean precision, mean recall and F-score ratings for comments on 4 products. The mean precision for the external hard drive was higher than the rest, and 3 and 4 rates were well predicted. The highest performance is in the various metrics except for the precision is for the mobile. The reason for the higher percentage of mobile was the number of comments first, compared with the other products, and secondly, the number of classes was 3, which resulted in a better prediction accuracy, with most of the comments being attributed to one of the three classes, which has led to two classes There is no other right sorting.

Figure 6 Final result of classification (P5)
In this process, given the accuracy of predictions, these scores need to be improved. Regarding these results, a number of users do not comment correctly and scrupulously. The accuracy of users in the ratings and reviews is very effective in the output. To be more precise, it's better to increase the rating periods, instead of 1 to 5 users have the ability to rate from 1 to 100. Also, if there are more comments on the various products like mobile, we will have more precise predictions, so there is a need to encourage users to more comment the experience of working with products to increase comments and to have a more accurate classification.

**Conclusion**

As mentioned, online shopping has a greater advantage than physical shopping stores, so competition among online shops is very high. Considering that customers play an important role in these businesses, the need for customer relationship management is increasingly felt. User comments need to be taken into consideration in order to keep past customers and attract new customers, which in our research we analyzed reviews from digikala website. At first, various product groups were selected for product selection for analysis. The result of the observation was digital products for having better feedbacks. To improve this with advertising, trust, and better service, you can increase the number of non-digital consumer comments, which will increase the number of buyers by reading comments in the face of high-quality services. In the following, four products were selected: External hard, smartphone, speaker and tablet from the Digital Products section, then extracted, translated and stored in Excel format. Analysis on comments was created in the form of processes in the RM software. This study can be used to support management decisions in web design, products introductions, and to utilize the suppliers in the customers' interests. Below are some suggestions for future research in this area:

- Using Python for Persian Text mining by the hazm library and using stemming techniques in Persian language
- Text mining and sentiment analyzing of tweets containing the hashtag of Digikala
- Use more products with the same categories as well as analysis on non-digital products
- Sentiment Analysis for different product categories

**References**


Comparative Analysis of Predictive Models for the Likelihood of Infertility in Women Using Supervised Machine Learning Techniques

Jeremiah Ademola Balogun, Ngozi Chidozie Egejuru, and Peter Adebayo Idowu

Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria

paidowu@oauife.edu.ng

Abstract

Infertility is a worldwide problem, affecting 8% – 15% of the couples in their reproductive age. WHO estimates that there are 60 - 80 million infertile couples worldwide with the highest incidence in some regions of Sub-Saharan Africa also infertility rate may reach 50% compared to 20% in Eastern Mediterranean Region and 11% in the developed world. Infertility has caused considerable social, emotional and psychological stress between couples, among families, within the individual concerned and the society at large. Historical data constituting information describing the risk factors of infertility alongside the respective infertility likelihood status of women was collected from Obafemi Awolowo University Teaching Hospital Complex (OAUTHC). The predictive model was formulated using naive Bayes’, decision trees and multi-layer perceptron algorithm – supervised machine learning algorithms. The formulated model was simulated using the Waikato Environment for Knowledge Analysis (WEKA) environment. The results of the performance evaluation of the machine learning algorithms showed that the C4.5 decision trees and the multi-layer perceptron with an accuracy of 74.4% each outperformed the naive Bayes’ algorithm. In addition, the decision trees algorithm recognized variables relevant to predicting infertility and a rule that can be applied on patient risk factor records for infertility likelihood prediction was deduced from the tree structure. This showed how effective machine learning algorithms can be used in predicting the likelihood of infertility in Nigerian women.

Keywords: prediction model, infertility in women, multi-layer perceptron, decision trees, naïve bayes.

Introduction

While there is no universal definition of infertility, a couple is generally considered clinically infertile when pregnancy has not occurred after at least twelve months of regular sexual activity without the use of contraceptives [1]. Primary infertility is defined as childlessness and secondary infertility as the inability to have an additional live birth for a parous woman. Although women’s infertility is of greater research consideration, health care attention and social blame, male conditions cause or contribute to around half of all cases of infertility [2]. According to World Health Organization, infertility is defined as one year of frequent, unprotected intercourse during which pregnancy has not occurred [3]. In another definition, infertility is the inability of a sexually active woman who is not practicing contraception to have a live birth [4].

Early exposures (e.g. in utero or in childhood) could permanently reprogram men and women for fecundity or biologic capacity (e.g. gynecologic and urologic health or gravid health during pregnancy) and fertility outcomes (e.g. multiple births or gestational age at delivery), which could affect later adult on set diseases [5]. Thus, infertility could have public health implications beyond simply the inability to have children. Infertility can be attributed to any abnormality in the female or male reproductive system [3]. The etiology is mostly distributed fairly equally among the male and female with factors ranging from ovarian dysfunction, tubal factors amongst others. A smaller percentage of cases are attributed to endometriosis, uterine or cervical factors, or other causes. In approximately, one fourth of couples, the cause is uncertain and is referred to as unexplained infertility, while etiology is multifactorial for some couples [6].

In general, an infertility evaluation is initiated after 12 months of unprotected intercourse during which pregnancy has not been achieved. Earlier investigation may be considered when historical factors, such as previous pelvic inflammatory disease or amenorrhea suggest infertility, although physicians should be aware
that earlier evaluation may lead to unnecessary testing and treatment in some cases. Evaluation also may be initiated earlier when the female partner is older than 35 years, because fertility rates decrease and spontaneous miscarriage and chromosomal abnormality rates increase with advancing maternal age [7]. Partners should be evaluated together and separately, because each person may want to reveal information about which their partner is unaware, such as previous pregnancy or sexually transmitted disease.

The risk factors for infertility can be classified into: genital, endocrinial, developmental and general factors. Pelvic inflammatory disease (PID) due to sexually transmitted diseases, unsafe abortion, or puerperal infection is the main cause of tubal infertility caused mainly by chlamydial infection. Polycystic ovarian syndrome (PCOS) is thought to be the commonest cause of an ovulatory infertility [8]. Several lifestyle factors may affect reproduction, including habits of diet, clothing, exercise, and the use of alcohol, tobacco and recreational drugs. Exposure to textile dyes, lead, mercury and cadmium, volatile organic solvents and pesticides has been also associated with infertility [9]. Estimates of the proportion of infertility cases attributable to male or female specific factors in developed countries were derived in the 1980s by the WHO: 8% of infertility cases were attributable to male factors, 37% to female factors, 35% to both the male and female, and 5% to an unknown cause (the remaining 15% became pregnant) [10].

Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand, description focuses on finding patterns describing the data that can be interpreted by humans. Machine learning plays an important role in disease prediction by identifying related pattern that exists between the risk factors associated with the likelihood of infertility in women. This will improve the level of decision-support offered to the expert gynecologist during the course of diagnosis.

This study presents a comparative analysis between three (3) supervised machine learning model used to develop predictive models for the likelihood of infertility in women in order to propose the most effective and efficient model. Where possible, variables that are relevant to predicting the likelihood of infertility in women alongside their underlying relationship will also be proposed.

**Related Works**

There are different types of diseases whose likelihood or survival had been predicted using data mining technique namely Hepatitis and other liver disorders, Breast cancer, Thyroid disease, Diabetes, HIV/AIDS and Tuberculosis etc., for the purpose of this research, the prediction of likelihood of infertility in women, research work that are related to fertility were reviewed. There existed a number of research areas concerning infertility but none attempts to predict its likelihood in women using data mining technique, further to its prediction is the usage of a graphical user interface or rather a software system.

Durairaj and Kumar [11] worked on Selection of Influential Parameters on Fertility using a data mining method of data analysis, as classification is proposed for the In-Vitro Fertilization (IVF) data analysis, and multilayer perceptron network for classification or prediction. From the experiments, the observation was made in the attribute selection analysis and it helped to identify the most influential IVF parameters to predict the successful rate of IVF treatment. The proposed technique was useful for finding the minimum set of influential parameters in order to predict a success rate of IVF, which enabled the gynecologists to prescribe the treatment to the couples. By knowing the success rate prior to the treatment, the couples get psychological boost, which increases their chances of getting successful pregnancy.

Saith et al. [12] used decision trees to investigate the relationship of the features of the embryo, oocyte and follicle to the successful outcome of the embryo transfer. Although 53 features were studied, only 4 had predictive capabilities, embryo grade, cell number, follicle size and follicular fluid volume. This study used 200 IVF records and significantly differs from our study in that it did not consider any clinical data on the female and male patients involved in the procedure.
Shen et al [13] used statistical analysis to examine factors involved in IVF procedures. This study, however, only considered fertilizations accomplished with Intracytoplasmic Sperm Injection (ICSI). Statistical approaches were used to find that sperm motility and ICSI operator were the two most important predictors for the success of an IVF procedure. Sperm motility and ICSI technician were also features considered in the study. The data set was drastically different because the ICSI method of fertilization was used in only 44 % of their records.

Methods

Data Collection

For the purpose of this study, it was necessary to identify and collect the data needed for identifying infertility in women from gynecologist located at the Obafemi Awolowo University Teaching Hospital Complex (OAUTHC) and the Faculty of Health Sciences of Obafemi Awolowo University, Ile-Ife. The variables identified include: age of menarche, age of marriage, family history of infertility, menstrual cycle, diabetes mellitus, hypertension, thyroid disease, pelvi-abdominal operation, endometriosis, fibroid disease, polycystic ovary, genital infection, previous termination of pregnancy, Sexually Transmitted Infection (STI) and the likelihood of infertility (identified using the labels: Likely, Unlikely and Probably) (Table 1). Data was collected from a total of 39 patients with a description of the variables in the dataset stated as follows:

a. Age of Menarche: is the identification of the age of the patient at first menstruation; it is recorded as a nominal value which determines the age category in years identified as equal or less than 15 years and greater than 15 years.

b. Age of marriage: is the identification of the patient’s age of marriage; it is recorded as a nominal value less than or equal to 30 years and greater than 30 years.

c. Menstrual cycle: is the identification of the regularity of the patient’s menstrual cycle; it is a nominal value identified as Regular or Irregular.

d. Family history of Infertility: is the identification of an existing history of infertility in the family; it is a nominal value identified as either Yes or No.

Table 1: Identified variables for determining infertility

<table>
<thead>
<tr>
<th>S/N</th>
<th>Class of Risk</th>
<th>Risk Factors/Considered Parameters (Points)</th>
<th>Labels (Points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Personal Profiles</td>
<td>Age of Menarche</td>
<td>≤15 yrs or &gt;15 yrs</td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td>Age of Marriage</td>
<td>≤30 yrs or &gt;30 yrs</td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td>Family History of infertility</td>
<td>Yes or No</td>
</tr>
<tr>
<td>4.</td>
<td></td>
<td>Menstrual cycle</td>
<td>Regular or Irregular</td>
</tr>
<tr>
<td>5.</td>
<td>Medical and Surgical history</td>
<td>Diabetes Mellitus</td>
<td>Yes or No</td>
</tr>
<tr>
<td>6.</td>
<td></td>
<td>Hypertension</td>
<td>Yes or No</td>
</tr>
<tr>
<td>7.</td>
<td></td>
<td>Thyroid</td>
<td>Yes or No</td>
</tr>
<tr>
<td>8.</td>
<td></td>
<td>Pelvi-abdominal operation had</td>
<td>Yes or No</td>
</tr>
<tr>
<td>9.</td>
<td>Gynecological history</td>
<td>Endometriosis</td>
<td>No or Yes</td>
</tr>
</tbody>
</table>
10. Fibroid
11. Polycystic Ovary
12. Genital Infection
13. Sexually transmitted Infection (STI)
14. Previous termination of pregnancy

Data-Preprocessing

Following the collection of data from the required respondents; 39 patients with their respective attributes (14 infertility risk indicators) alongside the likelihood of infertility was identified. In addition, the task of data cleaning for noise removal (errors, misspellings etc.) and missing data were performed on the information collected from the health records. Following this process, all data cells describing the attributes (fields) of each patient were found to be filled. No missing data were found in the repository and all misspellings were corrected.

In order for the dataset collected to be fit for the simulation environment; the dataset was converted to a more compactible data storage format. This would make the dataset fit for all the necessary machine learning operations performed by the simulation environment. Important to the study is the ability of the machine learning techniques to identify the most important combination of features that are more likely to improve the predicting the likelihood of infertility.
The dataset collected was converted to the required format needed for simulation; the Waikato Environment for Knowledge Analysis (WEKA) called the attribute relation file format (.arff) – a light-weight java application with a number of supervised and unsupervised machine learning tools. This format allows for the formal identification of the file name, attribute names and labels alongside the dataset that correspond to each attribute expressed using their respective labels. Figure 1 shows the format of the .arff file format chosen for the formal representation of the dataset using the 39 patient data collected.

![Figure 1: arff file containing identified attributes](image)

**Model Formulation**

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. Supervised machine learning algorithms make it possible to assign a set of records (infertility risk indicators) to a target classes – the risk of infertility (Unlikely, Likely and Benign).

Supervised machine learning algorithms are Black-boxed models, thus it is not possible to give an exact description of the mathematical relationship existing among the independent variables (input variables) with respect to the target variable (output variable – risk of infertility). Cost functions are used by supervised machine learning algorithms to estimate the error in prediction during the training of data for model development. Gradient decent and other related algorithms are used to reduce the error by estimating cost function parameters.

**Naïve Baye’s Classifier**

Naïve Bayes Classifier is a probabilistic model based on Baye’s theorem. It is defined as a statistical classifier. It is one of the frequently used methods for supervised learning. It provides an efficient way of handling any number of attributes or classes which is purely based on probabilistic theory. Bayesian classification provides practical learning algorithms and prior knowledge on observed data.
If X is a data sample containing instances, X_i where each instances are the infertility likelihood risk factors. Let H be a hypothesis that X belongs to class C which contains likely, probable and unlikely cases. Classification requires the determination of the following:

- \( P(H_j|X) \) – the posteriori probability: the probability that the hypothesis, Hj (unlikely, benign or likely) holds given the observed data sample X.
- \( P(H_j) \) - prior probability: the initial probability of the class, j;
- \( P(X_i) \): probability that sample data is observed for each attribute, i;
- \( P(X_i|H) \) - likelihood: the probability of observing the sample's attribute, X_i given that the hypothesis holds in the training data X; and

The posteriori probability of a hypothesis Hj defined as either of unlikely, likely or benign, \( P(H_j|X) \), follows the Baye’s theorem as follows:

\[
P(H_j|X) = \prod_{i=1}^{n} \frac{P(X_i|H_j)P(X_i)}{P(H_j)} \quad \text{for} \quad j = 1,2,3
\]

Where \( X = \{X_1, X_2, X_3 \ldots \ldots \ldots X_n\} \) is the set of risk factors for infertility likelihood of each patient, X and \( H_j = \{H_1 = \text{likely}, H_2 = \text{probable}, H_3 = \text{unlikely}\} \) is the target class set.

The breast cancer risk output class is thus:

\[
\max \{P(H_j|X)\} \quad \text{for} \quad j = 1,2,3.
\]

**Decision Trees Algorithm**

The theory of a decision tree has the following parts: a root node is the starting point of the tree; branches connect nodes showing the flow from question to answer. Nodes that have child nodes are called interior nodes. Leaf or terminal nodes are nodes that do not have child nodes and represent a possible value of target variable given the variables represented by the path from the root. The rules are inducted by definition from each respective node to branch to leaf.\(^\text{14}\)

Splitting points attribute variables and values of chosen variables are chosen based on Gini impurity (eqn. 3) and Gini gain (eqn. 4) as expressed below by Chaurasia et al.\(^\text{14}\):

\[
i(t) = 1 - \sum_{i=1}^{m} f(t,i)^2 = \sum_{i \neq j} f(t,i)f(t,j) \quad (3)
\]

\[
\Delta i(s,t) = i(t) - P_L \cdot i(t_L) - P_R \cdot i(t_R) \quad (4)
\]

Where \( f(t,i) \) is the probability of getting i in node t, and the target variable takes values in \( \{1, 2, 3 \ldots \text{m} \} \). \( P_L \) is the proportion of cases in node t divided to the left child node and \( P_R \) is the proportion of cases in t sent to the right child node. If the target variable is continuous, the split criterion is used with the Least Squares Deviation (LSD) as impurity measure. If there is no Gini gain or the preset stopping rule are satisfied, the splitting process stops.

Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S.
• Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition $S$ into corresponding subsets $S_1, S_2, ...$ according to the outcome for each case, and apply the same procedure recursively to each subset.

ID3 (Iterative Dichotomiser 3) developed by Ross Quinlan$^{15}$ is a classification tree used in the concept of information entropy. This provides a method to measure the number of bits each attribute can provide, and the attribute that yields the most information gain becomes the most important attribute and it should go at the top of the tree. Repeat this procedure until all the instances in the node are in the same category.

In this study, there are three outcomes, namely: Likely ($u_1$), Unlikely ($u_2$) and probably ($u_3$) in the root node $T$ of target variable. Let $u_1, u_2$ and $u_3$ denote the number of probable, unlikely and likely records, respectively. The initial information entropy is given by equation 5 as:

$$I(u_1, u_2, u_3) = - \sum_{i=1}^{3} \frac{u_i}{u_1 + u_2 + u_3} \log_2 \frac{u_i}{u_1 + u_2 + u_3}$$

(5)

If attribute $X$ (a risk indicator of infertility) with values $\{x_1$ and $x_2\}$ is chosen to be the split predictor and partition the initial node into $\{T_1, T_2, T_3... T_N\}$, and $u_1, u_2$ and $u_3$ denote the number of probable, unlikely and likely records in the child node $j$. The expected information entropy, $EI(X)$ and information gain, $G(X)$ are given by:

$$EI(X) = \sum_{j=1}^{N} \frac{u_{1j} + u_{2j} + u_{3j}}{u_1 + u_2 + u_3} \cdot I(u_1, u_2, u_3),$$

(6)

$$G(X) = I(u_1, u_2, u_3) - EI(X)$$

(7)

In 1993, Ross Quinlan made several improvements to ID3 and extended it to C4.5$^{15}$. Unlike ID3 which deals with discrete attributes, C4.5 handles both continuous and discrete attributes by creating a threshold to split the attribute into two groups, those above the threshold and those that are up to and including the threshold. C4.5 also deals with records that have unknown attribute values. C4.5 algorithm used normalized information gain or gain ratio as a modified splitting criterion of information gain which is the ratio of information gain divided by the information due to the split of a node on the basis of the value of a specific attribute. The reason of this modification is that the information gain tends to favor attributes that have a large number of values.

**Multi-layer Perceptron Architecture**

Multi-layer perception (MLP) is a natural extension of the single layer perception network of the class of artificial neural networks used in artificial intelligence. It is characterized by a forward flow of a set of inputs passing through subsequent hidden and computational layers composed by perception neurons using the feed-forward algorithm (Figure 3). The usage of MLPs is defended by the fact that they are able to predict and detect more complicated patterns in data. This is because multi-layer perceptron uses an additional algorithm which is called the back-propagation algorithm.
Figure 3: Structure of the multi-layer perceptron architecture

The back-propagation algorithm used in this study to train the network consists of two steps:

i. **Step 1 - Forward pass:** the inputs are passed through the network layer by layer and an output is produced. During this step, the synaptic weights are fixed; and

ii. **Step 2 - Backward pass:** the output from step 1 is compared to the target producing an error signal. That is propagated backwards. The aim of this step is to reduce the error in a statistical sense by adjusting the synaptic weights according to a defined scheme.

The multilayer perception has the following characteristics:

i. At all neurons within the network feature, a nonlinear activation function that is differentiable is present everywhere;

ii. The network has one or more hidden layers made up of neurons that are removed from direct contact with input and output. These neurons calculate a signal expressed as a nonlinear function of its input with synaptic weights and an estimate of the gradient vector; and

iii. There is a high degree of interconnectivity within the network.

The mathematical model of the multi-layer perceptron in Figure 3 is as follows:

- **The Input Layer**

In this part of the multi-layer perception (MLP) the input values, $X_1$ (factors responsible for infertility in women) are entered into the MLP system where $n$ is the number of attributes ($n=14$ in this study) and the weights, $W_i$ of each input, $X_i$ produce a summation, $U_k$ which is added to a bias variable, $X_0$ (takes a value of 0 or 1) all equal to $V_k$ is sent to the hidden layer for the activation function, $\phi$ to take effect where $k$ is the hidden layer. The Summation $U_k$ has the expression as follows:

$$U_k = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \cdots + w_nx_n$$  \hspace{1cm} (8a)

$$Thus, V_k = U_k + x_0$$ \hspace{1cm} (8b)

Where $X_n = (x_1, x_2, x_3, \cdots, x_n)$ is the patient’s record containing the factors considered predictive for the prediction of infertility in women.
And n = 14 attributes (input variables, \( x_n \))

- **The Hidden Layer**

At this part of the MLP the summation of the input variables are all sent to the activation function which is fired through all the hidden layers (for the purpose of this study 20 layers was used) using the activation function called the sigmoid function. The sigmoid function is expressed as:

\[
Sigmoid \ function, \ \varphi = \frac{1}{1 + e^{-av}} \tag{9}
\]

Where \( a \in \mathbb{R} \) is a shape parameter of the sigmoid function

And \( v = V_k \)

- **The Output Layer**

At this point, the value of the output (infertility status) is determined with the error rate as low as possible. Also, the back-propagation algorithm is applied which tries to reduce the error rate of the model via gradient descent by adjusting the values of the synaptic weights before the neuron fires the next set of inputs. At iteration \( m \) (the \( m \)th row in the training set) which in this case is 39, the error for neurons in the output layer is calculated in order to determine the error in computation. The error is calculated thus:

\[
error, \ \varepsilon = y_{pi} - y_{ai} \tag{10}
\]

Gradient descent is \( \lim_{t \to k} \frac{d\varepsilon}{dt} = 0 \) where \( t = k \) is the number of iterations

\[
\text{mean square error} = \frac{1}{2m} \cdot \sum_{i=1}^{m} (y_{pi} - y_{ai})^2 \tag{11}
\]

Where \( y_{pi} \) and \( y_{ai} \) are the predicted and actual output for patient, \( i \)

And \( m \) is the total patient data (\( m = 39 \))

**Performance Evaluation**

Following the development of the predictive model using all the proposed methods, the performance of the model was evaluated using the confusion matrix to determine the value of the performance metric chosen for this study. A confusion matrix contains information about actual and predicted classification done by a classification system and its performance is commonly evaluated using the data in the matrix (Figure 4). In this study, the likely cases are the positive cases while the probable and unlikely cases are the negative cases. Also, correctly classified cases are placed in the true cells (positive and negative) while incorrect classifications are placed in the false cells (positive and negative) and this has generated the rule (i) to (iv), below:

i. True positives (TP) are correctly classified positive cases;

ii. False positives (FP) are incorrectly classified positive cases;

iii. True negatives (TN) are correctly classified negative cases; and

iv. False negatives (FN) are incorrectly classified negative cases.
Figure 4: Diagram of a Confusion Matrix

From a confusion matrix, different measures of the performance of a prediction model can be determined using the values of the true positive/negatives and false positives/negatives. For the purpose of this study, the positive cases are the Likely Cases of infertility while the negative cases are probably and Unlikely cases.

a. True Positive rates (TP rates/Recall) – proportion of positive cases correctly classified

\[ TP \text{ rate} = \frac{TP}{TP + FN} \] (12)

b. False Positive rates (FP rates/False alarms) – proportion of negative cases incorrectly classified as positives

\[ FP \text{ rate} = \frac{FP}{FP + TN} \] (13)

c. Precision – proportion of predicted positive cases that were correct

\[ \text{Precision} = \frac{TP}{TP + FP} \] (14)

d. Accuracy – proportion of the total predictions that was correct.

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \] (15)

Results

Data Description

The data containing information about the attributes and the respective infertility status for 39 patients is shown in Table 2 alongside the distribution of the data shown in Figure 5. It was observed that out of the 39 patients, 19 were likely infertile, 3 were probably infertile and 17 were unlikely infertile. The highest distribution was: 23 with age of menacere less than or equal to 15 years, 23 had thyroid disease, 22 had no family history of infertility, 20 had no previous terminated pregnancy, 21 had irregular menstrual cycle, 21 had diabetes mellitus, 21 had hypertension, 21 had polycystic ovary and 21 had no genital infection.
The lowest distribution was: 16 had age of menacre more than 15 years, 16 had no thyroid disease, 17 had family history of infertility, 17 had previously terminated pregnancy, 18 had irregular menstrual cycle, 18 had no diabetes mellitus, 18 had no hypertension, 18 had no polycyclic ovary and 18 had genital infection.

### Table 2: Description of the identified variables

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Attributes</th>
<th>Labels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Age of Menacre</td>
<td>&lt;=15 years</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;15 years</td>
<td>16</td>
</tr>
<tr>
<td>K</td>
<td>Age of Marriage</td>
<td>&lt;=30 years</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;30 years</td>
<td>19</td>
</tr>
<tr>
<td>N</td>
<td>Family History of Infertility</td>
<td>No</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>17</td>
</tr>
<tr>
<td>P</td>
<td>Menstrual Cycle</td>
<td>Irregular</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Regular</td>
<td>18</td>
</tr>
<tr>
<td>U</td>
<td>Diabetes Mellitus</td>
<td>No</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>21</td>
</tr>
<tr>
<td>T</td>
<td>Hypertension</td>
<td>No</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>21</td>
</tr>
<tr>
<td>T</td>
<td>Thyroid Disease</td>
<td>No</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>23</td>
</tr>
<tr>
<td>T</td>
<td>Pelvi-Abdominal Operation</td>
<td>No</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>19</td>
</tr>
<tr>
<td>T</td>
<td>Endometriosis</td>
<td>No</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>20</td>
</tr>
<tr>
<td>T</td>
<td>Fibroid</td>
<td>No</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>19</td>
</tr>
<tr>
<td>T</td>
<td>Polycyclic Ovary</td>
<td>No</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>21</td>
</tr>
<tr>
<td>T</td>
<td>Genital Infection</td>
<td>No</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>18</td>
</tr>
<tr>
<td>T</td>
<td>Previous Terminated Pregnancy</td>
<td>No</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yes</td>
<td>17</td>
</tr>
<tr>
<td>OUTPUT</td>
<td>Infertility Status</td>
<td>Likely</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Probably</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unlikely</td>
<td>17</td>
</tr>
</tbody>
</table>

### Simulation Results

Three different supervised machine learning algorithms were used to formulate the predictive model for the likelihood of infertility; they were used to train the development of the prediction model using the dataset containing 39 patients' risk factor records. The simulation of the prediction models was done using the Waikato Environment for Knowledge Analysis (WEKA). The C4.5 decision trees algorithm was implemented using the J48 decision trees algorithm available in the trees class, the naïve Bayes' algorithm was implemented using the naïve Bayes' classifier available in the Bayes class while the Multi-layer perceptron was implemented using the multi-layer perceptron classifier available in the functions class all available on the WEKA environment of classification tools. The models were trained using the 10-fold cross validation method which splits the dataset into 10 subsets of data – while 9 parts are used for training the remaining one is used for testing; this process is repeated until the remaining 9 parts take their turn for testing the model.
Results of the naïve Bayes’ classifier

Using the naïve Bayes’ classifier to train the predictive model developed using the training data via the 10-fold cross validation method, it was discovered that there were 28 (71.79%) correct classifications and 11 (28.21%) incorrect classifications – showing an accuracy of 71.8% (Figure 5).

Using the confusion matrix, it was discovered that out of 19 likely cases there were 15 correct classifications while 1 misclassified for probable and 3 for unlikely. Out of 3 probable cases there were no correct classifications while 1 misclassified for likely and 2 for unlikely. Out of 17 unlikely cases there were 13 correct classifications with 3 misclassified for likely and 1 for probable (Figure 6 – left). Figure 7 shows a graphical plot of the correct and incorrect classifications – correct classifications are crosses while incorrect classifications are boxes.
Results of the C4.5 decision trees classifier

Using the C4.5 decision trees classifier to train the predictive model developed using the training data via the 10-fold cross validation method, it was discovered that there were 29 (74.36%) correct classifications and 10 (25.64%) incorrect classifications – showing an accuracy of 74.4% (Figure 8). Using the confusion matrix, it was discovered that out of 19 likely cases there were 15 correct classifications while 1 misclassified for probable and 3 for unlikely. Out of 3 probable cases there were no correct classifications while 2 misclassified for likely and 1 for unlikely. Out of 17 unlikely cases there were 14 correct classifications with 3 misclassified for likely (Figure 6 – middle). Figure 9 shows a graphical plot of the correct and incorrect classifications – correct classifications are crosses while incorrect classifications are boxes.
For every decision trees algorithm there is always a hierarchical tree with an attribute at each node form the parent node all the way to the child node to the leave - the target class. The tree can be converted to a rule by following the pattern from the parent node at the top all the way to the child node until the bottom leaf is achieved where the necessary classification is defined. Figure 10 shows the decision trees constructed during the model development; it can be seen that a number of variables were identified as been relevant for infertility likelihood prediction. It can also be discovered that the size of the tree is 6 and the number of leaves plotted are 5. The variables identified are:

- Previous termination of pregnancy
- Menstrual Cycle
- Age of Manacre and
- Genital Infection
Figure 10: Graphical plot of the decision tree for infertility likelihood

Using the decision tree in Figure 10, the following rule can be used to predict the likelihood of infertility in women given the values of the four identified risk factors. The rule can be read as follows:

IF Previous Termination of Pregnancy = “Yes” THEN infertility likelihood = “Likely”

Else IF Previous Termination of Pregnancy = “No” THEN

IF Menstrual Cycle = “Regular” THEN infertility likelihood = “Unlikely”

Else IF Menstrual Cycle = “Irregular” THEN

IF Age of Menacre = “>15 years” THEN infertility likelihood = “Probable”

Else IF Age of Menacre = “<=15 years” THEN

IF Genital Infection = “Yes” THEN infertility likelihood = “Likely”

Else IF Genital Infection = “No” THEN infertility likelihood = “Unlikely”

Results of the Multi-Layer Perceptron (MLP) classifier

Using the Multi-layer perceptron classifier to train the predictive model developed using the training data via the 10-fold cross validation method, it was discovered that there were 29 (74.36%) correct classifications and 10 (25.64%) incorrect classifications – showing an accuracy of 74.4% (Figure 11). Using the confusion matrix, it was discovered that out of 19 likely cases there were 16 correct classifications while 2 misclassified for probable and 1 for unlikely. Out of 3 probable cases there were no correct classifications while 1 misclassified for likely and 2 for unlikely. Out of 17 unlikely cases there were 13 correct classifications with 1 misclassified for likely and 3 for probable (Figure 6 – right). Figure 12 shows a graphical plot of the correct and incorrect classifications – correct classifications are crosses while incorrect classifications are boxes.
Discussions

Table 3 gives a summary of the simulation results by presenting the average value of each performance metrics that was evaluated for the machine learning techniques used. The True positive rate (recall/sensitivity), false positive rate (false alarm/1-specificity), precision, accuracy and the area under the receiver operating characteristics (ROC) curve were used. From the table, it was discovered that the decision trees and the MLP algorithms showed the highest accuracy due to the ability to predict 29 out of the 39 records correctly. The true positive rate was also highest for the decision trees and the MLP algorithms with an equal value of 0.744 – which implies that 74.4% of the actual positive cases (likely) were correctly classified. The MLP showed the lowest value for the false positive rate with a value of 0.119 – which implies that 11.9% of the actual negative classes (probable or unlikely) were misclassified for positive cases. The MLP also had the highest value for the precision with a value of 0.787 – which implies that 78.7% of the positive classifications made were actually positive classes. The decision trees algorithm was observed to have the lowest area under the receiver operating characteristics (ROC) curve – a graph of the TP rate against the FP rate which had a value of 0.722. The area under the graph is used to identify the level of relevance that can be given to the machine learning algorithm at making predictions – thus, the higher the value then the lower the bias of the model.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Accuracy (%)</th>
<th>TP rate (recall)</th>
<th>FP rate (False alarm)</th>
<th>Precision</th>
<th>Area under ROC Curve (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes’</td>
<td>71.795</td>
<td>0.718</td>
<td>0.201</td>
<td>0.699</td>
<td>0.855</td>
</tr>
<tr>
<td>Decision Trees</td>
<td><strong>74.359</strong></td>
<td><strong>0.744</strong></td>
<td>0.203</td>
<td>0.704</td>
<td>0.722</td>
</tr>
</tbody>
</table>
From the simulation results, it can be inferred that the most effective supervised machine learning algorithm is the multi-layer perceptron (MLP) due to its high accuracy, TP rate and Precision with lower value for the FP rate. The variables identified and the rule deduced from the variables using the decision trees algorithm can also be used to support decision made by gynecologist concerning infertility likelihood in women.

Conclusions

In this paper, the development of a predictive model for determining the likelihood of infertility in Nigerian women was proposed using dataset collected from patients in Obafemi Awolowo University Teaching Hospital Complex (OAUTHC), Ile-Ife, Osun State in Nigeria. 14 variables were identified by gynecologist to be necessary in predicting infertility in women for which a dataset containing information of 39 patients alongside their respective infertility status (likely, unlikely and probably) was also provided with 14 attributes following the identification of the required variables.

After the process of data collection and pre-processing, three supervised machine learning algorithms were used to develop the predictive model for the likelihood of infertility in women using the historical dataset from which the training and testing dataset was collected. The 10-fold cross validation method was used to train the predictive model developed using the machine learning algorithms and the performance of the models evaluated.

The multi-layer perceptron proved to be an effective algorithm for predicting infertility in women given the attributes identified but it is believed that higher accuracy could be attained by increasing the number of records used and be identifying other relevant attributes which could help predict infertility in women. Rule induced algorithms can also be used to plot the relationship between the selected attributes identified with respect to determining the likelihood of infertility in women using the decision trees algorithm.

References


An Ensemble Model of Machine Learning Algorithms for the Severity of Sickle Cell Disease (Scd)
Among Paediatrics

1Balogun Jeremiah Ademola, 2Aderounmu, Temilade, Egejuru Ngozi Chidozie and 1Idowu, Peter Adebayo

1Department of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.
2Department of Paediatrics and Child Health, Obafemi Awolowo University Teaching Hospital Complex (OAUTHC), Ile-Ife, Nigeria.

Abstract

This study was motivated at developing an ensemble of 3 supervised machine learning algorithms for the assessment of the severity of sickle cell disease (SCD) among paediatric patients. The study collected data from a tertiary hospital in south-western Nigeria following the identification of variables required for assessing the severity of SCD. The study also adopted the use of 3 supervised machine learning algorithms namely: naïve Bayes (NB), C4.5 decision trees (DT) and support vector machines (SVM) for creating the ensemble model using a 10-fold cross validation technique. The models were created by adopting the algorithms in isolation and in combination of 2 and 3 which were compared. The developed models were evaluated in order to present the model with the best performance. The results of the study showed that using an ensemble of DT and NB alone provided the best performance. The study has implications in presenting a model for improving the assessment of the severity of SCD among paediatric patients in Nigeria.

Keywords: Sickle Cell Disease (SCD), Disease severity, Stack-Ensemble Model, Naïve Bayes, Decision Trees, Multi-Layer Perceptron.

1. Introduction

Sickle cell disease (SCD) is a genetic blood disorder – a structural variant of normal haemoglobin which affects the red blood cells of humans and has led to high morbidity and mortality rates thereby becoming a global public health concern (Chakravorty & Williams, 2015). According to the World Health Organization (WHO, 2006), it was recommended that 50% of member states establish SCD control programs by the year 2020. Aliyu et al. (2008), reported that there are between 20 and 25 million people worldwide living with SCD among which 12 to 15 million live in Africa.

According to Agasa et al. (2010), the highest prevalence of sickle-cell trait (SCT) in Africa usually occur around tropical areas which lie between latitudes of 15° North and 20° south with SCD prevalence which range between 10 and 40 percent of the population. It was also estimated that 240,000 children were born with SCD annually in sub-Saharan Africa (Makani et al., 2011). It was also estimated that 75 to 85 percent of children born with SCD were born in Africa, where mortality rates for those under the age of 5 years range from 50 to 80 percent. In 2012, it was reported that SCD had affected 20 to 25 million people globally among which 50 to 80 percent of infants born with SCD in Africa die before the age of 5 years (Aygun & Odame, 2012).

Life expectancy in SCD was substantially reduced especially in those with severe disease as reported in a 10-year retrospective study which revealed that the mean age of SCD patients was found to reduce, suggesting reduced life expectancy. Anemia is a major cause of morbidity and mortality in SCD, and many patients die in hospital emergency rooms and wards before blood transfusions can be initiated (Ikefuna & Emodi, 2007). It has been suggested that one factor associated with the high incidence of SCD in tropical Africa is the protection against Plasmodium malaria associated with having the SCD (Aygun & Odame, 2012).

Machine learning (ML) is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization tools to learn from past examples and afterwards use the prior training to classify new data, identify new patterns or predict novel trends (Mitchell 1997). Machine Learning has also been extensively adopted in
medical research to generate knowledge from complex clinical data required for improving clinical decision making process (Jaree et al., 2013). Observational studies show that data mining and machine learning prediction techniques have been widely used to determine patterns and how these patterns can be used by physicians to determine diagnoses, prognosis and apply treatment for patients (Boris and Milan, 2012).

The adoption of machine learning into healthcare research has also shown success in the prediction and diagnosis of various diseases thus increasing the accuracy of diagnosis and provide answers to physicians about affected patients (Jiang et al., 2017). Ensemble learning refers to the procedures employed to train multiple learning machines and combine their outputs, treating them as a committee of decision makers (Joshi & Srivastava, 2014). The success of the ensemble approach depends on the diversity in the individual classifiers with respect to misclassified instances (Simidjievski et al., 2016). There exist numerous methods for model combination which includes: linear combiner, the product combiner, and the voting combiner are by far the most commonly used in practice.

In Nigeria today, the number of children with sickle cell disorder (SCD) is increasing with every recorded birth thus leading to an increase in the number of deaths associated with SCD. The number of deaths associated with SCD has been in part as a results of poor management of SCD patients by uneducated and ignorant parents leading to the number of emergency visits to the hospitals for treatment for anaemia and crisis episodes. Related studies in the area of SCD have been targeted at either monitoring the risk of SCD or at the survival of SCD with but not to the management of SCD.

A number of related works which have adopted the use of machine learning to the healthcare data management have shown that a single classifier may have varying performance over a variety of data which can be removed by combining more than one classifier. This challenge has paved way for the development of ensemble methods which combine one or more machine learning algorithms for the development of predictive models. There is a need for the development of an ensemble of machine learning algorithms aimed at improving the assessment of the severity of SCD among paediatric patients in Nigeria, hence this study.

2. Related Works

A number of study have been reviewed in this study, among which include the application of machine learning to healthcare research and the adoption of the ensemble of various machine learning algorithms for predictive modeling.

Xiao et al. (2018), worked on the development of a deep learning-based multi-model ensemble method for cancer prediction. The study applied deep learning to an ensemble approach that incorporated 5 different machine learning models by supplying informative gene data selected by differential gene expression analysis to five different classification models. The results revealed that the proposed method showed an average accuracy of 98% however was limited to the demonstration of the advantage of voting ensemble learning over traditional machine learning techniques.

Xu et al. (2017), applied neural networks to the classification of red blood cells among SCD patients. The study proposed a method for automating the high-throughput Red Blood Cell (RBC) shape classification using a neural network framework. The presented a feature extraction of the region of interest (ROI) from RBC images following which the images captured were normalized. A convolutional neural network based classification system was formulated using 7000 single RBC images via 5 fold cross validation collected from 8 SCD patients. The results showed that the model was able to classify RBC with an accuracy of 67.5%. The study was limited to the classification of red blood cells.

Goyal and Kaur (2016), worked on a survey of application of ensemble modeling for loan prediction. The study identified that there were various techniques of ensemble which include: bagging, stacking, boosting, voting and using bucket of models to mention a few. The results of the survey showed that the development of
ensemble models guarantee better forecasting, a more constant model, better results and error reduction. The study was limited to a survey of the application of ensemble modeling for improving model performance.

King (2015) applied ensemble learning methods using various machine learning algorithms to structured and unstructured data which were collected, pre-processed, analyzed and followed by model evaluation. The study developed an ensemble model for classifying profitable campaigns thereby maximizing overall campaign portfolio profits. The study adopted the use of 4 traditional classifiers and 4 ensemble learning techniques to build models for identifying pay-per-click campaigns. The results if the study showed that using an ensemble configuration produced the highest campaign portfolio profit. The study was limited to the application of ensemble modeling to marketing data.

Milton et al. (2014) performed the prediction of fetal hemoglobin in sickle cell anemia using an ensemble of genetic risk prediction models. The study developed a collection of 14 models with genetic risk score (GRS) composed of different numbers of single nucleotide polymorphisms (SNPs), and use the ensemble of these models to predict HbF in sickle cell anemia patients. The models were trained in 841 sickle cell anemia patients and were tested in three independent cohorts. The ensemble of 14 models explained 23.4% of the variability in HbF in the discovery cohort, while the correlation between predicted and observed HbF in the 3 independent cohorts ranged between 0.28 and 0.44.

Otaigbe (2013) performed a study on the prevalence of blood transfusion in sickle cell anemia patients in south-south Nigeria over a two year period. The study involved the collection of data from the files of patients seen in clinic or admitted in the Pediatrics Department of the University of Port Harcourt Teaching Hospital within 2 years. Of the 131 cases observed, 130 had genotype Hb SS and 1 had genotype Hb SC. The results of the study showed that 57% had received at least one blood transfusion with the commonest indication been severe anaemia. The study concluded that efforts must be made to reduce the frequency of blood transfusion by monitoring the level of hematocrit in SCD patients.

3. Materials and Methods

This section identified the material and methods that were adopted for the development of the ensemble model required for assessing the severity of SCD among paediatric patients receiving treatments. It consists of a sequence of methods which started with the identification and the collection of data containing the features alongside the target classes of SCD severity. The ensemble model was formulated for the severity of anemia based on the data collected using a combination of the C4.5 decision trees (DT), Naïve Bayes (NB) and Support Vector Machines (SVM) classifiers. The ensemble model of classifiers was simulated using the Waikato Environment for Knowledge Analysis (WEKA) followed by a performance evaluation of model required for validating the ensemble model required for assessing the severity of SCD patients.

3.1 Method of data identification and collection

Following the review of related works of literature in the body of knowledge of SCD and its severity, a number of variables required for determining the severity of SCD were also determined. The identified variables for assessing the severity of SCD among paediatrics were validated by the medical experts with more than 10 years’ experience was interviewed before the data was collected from the medical records office of the Wesley Guilds, Obafemi Awolowo University Teaching Hospital Complex (OAUTHC) in Ilesha, Osun State.

The data was collected from paediatric SCD patients aged below 15 years. Information about the aforementioned variables was collected and stored into electronic format from the information stored in the files located at the medical records department of Wesley Guilds OAUTHC, Ilesha, Nigeria. The data collected was used for the formulation of the predictive model for determining the severity of SCD among paediatrics. A description of the variables are presented as follows.
**a. Gender (sex) of the patient:** was used to identify the gender of the SCD patients which was recorded as a nominal value Male (M) or female (F);

**b. Age of patients:** was used to identify the present age of the which was recorded as a numeric value (measured in years);

**c. Age at diagnosis:** was used to identify the age at which the SCD patients was screened for the presence and identification of SCD which was recorded as a numeric value (measured in years);

**d. Ethnicity:** was used to identify the ethnic tribe to which an SCD patient belonged to. It was measured as a nominal variable with values: Yoruba, Hausa, Ibo and others;

**e. Religion of SCD patient:** was used to identify the religion of the SCD patient that is receiving treatment and was measured using a nominal variable with values: Yoruba, Hausa, Ibo and others;

**f. Body Mass Index (BMI):** was used to identify the nutritional status of an SCD patients based on the values of the weight (measured in Kg) and the height (measured in meters). The BMI is a numeric values measured in Kg/m\(^2\) and can also be classified into nominal values such as: underweight, normal, obese and overweight;

**g. Clinical variables for assessing anemia risk:** These are a class of other variables which will be identified by the physician interviewed and will be used to estimate the risk of anemia among SCD patients receiving treatment. Such includes: the packet cell volume (PCV), frequency of anemia crisis, frequency of blood transfusions and so on.

### 3.2 Formulation of ensemble model of machine learning algorithms

This study adopted the development of an ensemble of three (3) machine learning algorithms which was formulated based on historical data collected from SCD patients. Figure 1 shows a diagram of how the ensemble model combined the 3 classifiers selected for this study. The ensemble model was formulated following the standard process of dividing the dataset into two (2) parts namely training and testing which adopted the 10-fold cross validation training technique. Therefore, the training dataset was used to formulate the models following which the testing dataset was used to validate the model.

**Figure 1: Ensemble Model for Anaemia Risk**

Equation (1) shows the mapping function that describes the relationship between the causative features and the target class used to assess the severity of SCD using the ensemble model \( \varphi \). The equation shows the relationship between the set of causative features represented by a vector, \( X \) consisting of the values of \( i \) variables and the label \( Y \) which defines the severity of SCD identified as Low, Moderate and High. Assuming the values of the set of variable for a SCD patient is represented as \( X = \{X_1, X_2, X_3, \ldots, X_i\} \) where \( X_i \) is the value of each variable, \( i \)
$\varphi: X \rightarrow Y$ 
defined as: $\varphi(X) = Y$  

$\varphi(X) = \begin{cases} 
\text{Low} \\
\text{Moderate} \\
\text{High} 
\end{cases}$

This study adopted a process of developing the predictive model for the severity of SCD using naïve Bayes’ (NB), support vector machines (SVM) and the C4.5 decision trees (DT) algorithms in isolation following which the ensemble model which combined the classifiers using a voting technique was formulated. Therefore, the first ensemble named Ensemble I combined DT and NB followed by Ensemble II which combined DT and SVM and Ensemble II which combined NB and SVM. Also, an ensemble model which combined all 3 classifiers, namely: DT, NB and SVM was also formulated called Ensemble IV. Following the formulation of the ensemble model by combining the respective classifiers, the testing dataset was used to validate the performance of the predictive model for the risk of anemia using a number of performance evaluation metrics. The algorithms adopted are presented in the following paragraphs.

- **C4.5 Decision Trees (DT) classifier**

The C4.5 decision trees classifier represents model evaluated from dataset as a hierarchical tree structure using a splitting criteria called the gain ratio. During the training process of model development using the historical dataset collected, the pattern was learned by the tree by splitting the training dataset into subsets based on an attribute value test for each input variables; the process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion was completed when the subset at a node had all the same value of the target class, or when splitting no longer adds value to the predictions. The criteria used by the C4.5 decision trees for tree split is the gain ratio which required the use of the information gain in equation (3) and split criteria in equation (4) to determine the gain ratio by dividing equation (3) by equation (4).

\[ IG(X_i) = H(X_i) - \sum_{t \in T} \frac{|t|}{|X_{ij}|} \cdot H(t) \]  

Where:

\[ H(X_i) = - \sum_{t \in T} \frac{|t \cdot X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t \cdot X_i|}{|X_{ij}|} \]

\[ Split(T) = - \sum_{t \in T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \] 

- **Naïve Bayes’ (NB) classifier**

Naïve Bayes’ Classifier is a probabilistic model based on Bayes’ theorem. It is defined as a statistical classifier. Bayesian classification provides practical learning algorithms and prior knowledge on observed data. Let $X_{ij}$ be a dataset sample containing records (or instances) of $i$ number of risks factors (attributes/features) alongside their respective severity of SCD, $C$ (target class) collected for $j$ number of records/patients and $H_k = \{ H_1 = \text{Low Risk}, H_2 = \text{Moderate Risk}, H_3 = \text{High Risk} \}$ be a hypothesis that $X_{ij}$ belongs to class $C$. For the classification of the risk of anaemia given the values of the risk factor of the jth record, Naïve Bayes’ classification required the determination of the following:
• \(P(H_k|X_{ij})\) – Posteriori probability: is the probability that the hypothesis, \(H_k\) holds given the observed data sample \(X_{ij}\) for \(1 \leq k \leq 3\).
• \(P(H_k)\) - Prior probability: is the initial probability of the target class \(1 \leq k \leq 3\);
• \(P(X_{ij})\) is the probability that the sample data is observed for each risk factor (or attribute), \(i\) and
• \(P(|X_{ij}|H_k)\) is the probability of observing the sample’s attribute, \(X_i\) given that the hypothesis holds in the training data \(X_{ij}\).

Therefore, the posteriori probability of a hypothesis \(H_k\) is defined according to Bayes’ theorem as follows:

\[
P(H_k|X_{ij}) = \frac{\prod_{i=1}^{n} P(X_{ij}|H_k)P(H_k)}{P(X_{ij})} \quad \text{for } k = 1,2,3
\]

Hence, the severity of SCD for a record is thus:

\[
\max [P(H_1|X_{ij}), P(H_2|X_{ij}), P(H_3|X_{ij})]
\]

• **Support vector Machines (SVM)**

An SVM model is a representation of the examples (data records) as points in space which were mapped so that the examples of the separate categories: Low, Moderate and High Risk were divided by a clear gap that is as wide as possible. In formal terms, the SVM was used to construct a hyper-plane in a high-dimensional space, and adopted for classification using the sequential minimum optimization (SMO) algorithms. A good separation was achieved by the hyperplane \(<w, x> + b = 0\) that has the largest distance \(\frac{2}{|w|}\) to the neighbouring data points of either classes at opposite ends, since in general the larger the margin the lower the generalization error of the SVM classifier. A hyperplane created is defined as \(<w, x> + b = 0\) where \(w \in \mathbb{R}^p\) and \(b \in \mathbb{R}\) while \(<w, x> + b = -1\) and \(<w, x> + b = 1\) are the margins required for the separation \(w\) of support vectors \(x\) within the \(n\) variables.

### 3.3 Simulation of ensemble model of machine learning algorithms

The Waikato Environment for Knowledge Analysis (WEKA) software – a suite of machine learning algorithms was used as the simulation environment for the development of the predictive model following the collection of data about paediatric SCD patients. The dataset collected was divided into two parts: training and testing data – the training data was used to formulate the model while the test data was used to validate the model. The process of training and testing predictive model according to literature is a very difficult experience especially with the various available validation procedures.

For this classification problem, it was required to measure a classifier’s performance in terms of the error rate. In order to predict the performance of a classifier on new data, there was the need to assess the error rate of the predictive model on a dataset that played no part in the formation of the classifier. This independent dataset was called the test dataset – which was a representative sample of the underlying problem as was the training data using the 10-fold cross validation technique.

The 10-fold cross validation technique involved the process of leaving a part of a whole dataset as testing data while the rest is used for training the model is called the holdout method. It involved dividing the whole datasets into 10 folds (or partitions) such that each partition was selected for testing with the remaining 9 partitions used for training. Each new partition was used for testing with the remaining successive 9 partitions (including the first partition used or testing) used for training until all 10 partitions had been selected for testing.

### 3.4 Validation of ensemble of machine learning algorithms

During the course of evaluating the predictive model, a number of metrics were used to quantify the model’s performance for model validation following model simulation using WEKA. These results of the correct and
incorrect classifications made by the ensemble model on the testing dataset was presented on a confusion matrix. For this study, the confusion matrix is a 3 x 3 confusion matrix table owing to the three (3) labels of the output class as shown in Figure 2. Using the confusion matrix, correct classifications were plotted along the diagonal from the north-west position for Low risk predicted as Low risk (A), followed by Moderate risk predicted as Moderate risk (E) and High risk predicted as High risk (I) on the south-east corner.

The incorrect classifications were plotted in the remaining cells of the confusion matrix. Also, the actual Low risk cases are A+B+C, actual Moderate risk cases are D+E+F, while actual High risk cases are G+H+I and the predicted Low risk cases are A+D+G, predicted Moderate risk cases are D+E+F and predicted High risk cases are G+H+I. The developed model was validated a number of performance metrics based on the values of A – I in the confusion matrix for each predictive model.

<table>
<thead>
<tr>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
<th>Predicted as</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>Low</td>
</tr>
<tr>
<td>D</td>
<td>E</td>
<td>F</td>
<td>Moderate</td>
</tr>
<tr>
<td>G</td>
<td>H</td>
<td>I</td>
<td>High</td>
</tr>
</tbody>
</table>

Figure 3: Confusion Matrix for Model Performance Evaluation

a. **Accuracy**: the total number of correct classification

\[
Accuracy = \frac{A + E + I}{A + B + C + D + E + F + G + H + I} \quad (7)
\]

b. **True positive rate (recall/sensitivity)**: the proportion of actual cases correctly classified

\[
TP_{Low} = \frac{A}{A + B + C} \quad (8a)
\]
\[
TP_{Moderate} = \frac{E}{D + E + F} \quad (8b)
\]
\[
TP_{High} = \frac{I}{G + H + I} \quad (8c)
\]

c. **False positive (false alarm/1-specificity)**: the proportion of negative cases incorrectly classified as positive

\[
FP_{Low} = \frac{B + C}{D + E + F + G + H + I} \quad (9a)
\]
\[
FP_{Moderate} = \frac{D + F}{A + B + C + G + H + I} \quad (9b)
\]
\[
FP_{High} = \frac{G + H}{A + B + C + D + E + F} \quad (9c)
\]

d. **Precision**: the proportion of predictions that are correct

\[
Precision_{Low} = \frac{A}{A + D + G} \quad (10a)
\]
Using the aforementioned performance metrics, the performance of the predictive model for the classification of risk of anemia can be evaluated by validation using a historical dataset collected based on the information provided in the questionnaire. The TP rate and precision lie within the interval [0, 1], accuracy within the interval of [0, 100] % while the FP rate lies within an interval of [0, 1]. The closer the accuracy is to 100% the better the model, the closer the value of the TP rate and precision is to 1 the better while the closer the value of FP rate is to 0 the better. Therefore, the evaluation of an effective model has a high TP/Precision rates and a low FP rates.

4. Results

This section presents the results of the methods adopted for the development of the ensemble model for the severity of SCD using 3 machine learning algorithms. The results of the description of the nominal and numeric attributes within the dataset were also identified. Following the presentation of the results of the description of attributes identified in this study for the development of the classification model for the severity of anemia among pediatric SCD patients the presentation of the results of the different models formulated based on the individual and ensemble of selected classifiers. The results of the evaluation of the performance of the predictive model was done based on the outcome of the testing phase using accuracy, true positive (TP) rate, false positive (FP) rate and precision with the model with the best performance presented.

4.1 Results of data identification and collection

The data considered in this study which was collected from South-west Nigeria contained demographic and clinical information about pediatric SCD patients. Based on the data collected from the patients, the severity of anemia among SCD patients was measured and recorded by the experts. The results of the data collection process showed that, majority of the SCD patients had Moderate severity owing for a proportion of 55.7% followed by SCD patients that had Low severity with a proportion of 33.9%.

Table 1 gives a description of the results of the data collected about SCD patients in terms of the severity of anemia among pediatric SCD patients. Table 2 shows a summary of the description of the nominal attributes among the identified attributes using a frequency distribution table in terms of the frequency of distribution and percentage of total. Based on the data presented in Table 2, it was observed that majority of the data collected consisted of male patients which was about 65% owing for a ratio of about 2 to 1 for male to female SCD patients.

Table 1: Results of the Distribution of the Severity of Anemia

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frequency</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Risk</td>
<td>39</td>
<td>33.9</td>
</tr>
<tr>
<td>Moderate Risk</td>
<td>64</td>
<td>55.7</td>
</tr>
<tr>
<td>High Risk</td>
<td>12</td>
<td>10.4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>115</strong></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>

The results of the data collected about the occupation of the parents of SCD patients showed that majority of the mothers were traders owing for a proportion of 46% followed by either artisans of teachers/civil servants with a proportion of 23.55 each. Regarding the occupation of the fathers, it the results showed that majority of
the fathers were either artisans or teacher/civil servants with a proportion of 32% each. The results regarding
the parent’s education based on the results showed that at least 30% of the parents had university education.
The results showed that majority of the parents were of upper class owing for a proportion of 40% followed by
middle class parents with proportion of 34%.

Table 2: Results of the Description of Nominal Attributes

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Values</th>
<th>Score</th>
<th>Frequency</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>1</td>
<td>75</td>
<td>65.2</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>2</td>
<td>40</td>
<td>34.8</td>
</tr>
<tr>
<td>Mother’s Education</td>
<td>No formal education</td>
<td>1</td>
<td>6</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>Primary</td>
<td>2</td>
<td>16</td>
<td>13.9</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>3</td>
<td>54</td>
<td>47.0</td>
</tr>
<tr>
<td></td>
<td>University</td>
<td>4</td>
<td>39</td>
<td>33.9</td>
</tr>
<tr>
<td>Mother’s Occupation</td>
<td>Full housewife</td>
<td>1</td>
<td>4</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>Artisan</td>
<td>2</td>
<td>27</td>
<td>23.5</td>
</tr>
<tr>
<td></td>
<td>Teacher/civil servant</td>
<td>3</td>
<td>27</td>
<td>23.5</td>
</tr>
<tr>
<td></td>
<td>Large scale business</td>
<td>4</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Professional</td>
<td>5</td>
<td>2</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td>6</td>
<td>2</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>Trader</td>
<td>7</td>
<td>53</td>
<td>46.1</td>
</tr>
<tr>
<td>Father’s Education</td>
<td>No formal education</td>
<td>1</td>
<td>5</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>Primary</td>
<td>2</td>
<td>6</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>Secondary</td>
<td>3</td>
<td>52</td>
<td>45.2</td>
</tr>
<tr>
<td></td>
<td>University</td>
<td>4</td>
<td>52</td>
<td>45.2</td>
</tr>
<tr>
<td>Father’s &amp; Occupation</td>
<td>Student</td>
<td>1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Artisan</td>
<td>2</td>
<td>37</td>
<td>32.2</td>
</tr>
<tr>
<td></td>
<td>Teacher/civil servant</td>
<td>3</td>
<td>36</td>
<td>31.3</td>
</tr>
<tr>
<td></td>
<td>Large scale business</td>
<td>4</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>Professionals</td>
<td>5</td>
<td>13</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td>Trader</td>
<td>6</td>
<td>17</td>
<td>14.8</td>
</tr>
<tr>
<td></td>
<td>Driver</td>
<td>7</td>
<td>11</td>
<td>9.6</td>
</tr>
<tr>
<td>Social Class</td>
<td>Upper Class</td>
<td>1</td>
<td>20</td>
<td>17.4</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>---</td>
<td>----</td>
<td>------</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>26</td>
<td></td>
<td>22.6</td>
</tr>
<tr>
<td>Middle Class</td>
<td>3</td>
<td>39</td>
<td></td>
<td>33.9</td>
</tr>
<tr>
<td>Lower Class</td>
<td>4</td>
<td>23</td>
<td></td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>7</td>
<td></td>
<td>6.1</td>
</tr>
<tr>
<td>CVD Lifetime Incidence</td>
<td>Yes</td>
<td>1</td>
<td>5</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2</td>
<td>110</td>
<td>95.7</td>
</tr>
<tr>
<td>Acute chest syndrome Lifetime Incidence</td>
<td>Yes</td>
<td>1</td>
<td>23</td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2</td>
<td>92</td>
<td>80.0</td>
</tr>
<tr>
<td>AVN Lifetime Incidence</td>
<td>Yes</td>
<td>1</td>
<td>5</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2</td>
<td>110</td>
<td>95.7</td>
</tr>
<tr>
<td>Pneumococcal Meningitis Lifetime Incidence</td>
<td>Yes</td>
<td>1</td>
<td>2</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2</td>
<td>113</td>
<td>98.3</td>
</tr>
<tr>
<td>Gall Stone Lifetime Incidence</td>
<td>Yes</td>
<td>1</td>
<td>3</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2</td>
<td>112</td>
<td>97.4</td>
</tr>
<tr>
<td>Osteomyelitis Lifetime Incidence</td>
<td>Yes</td>
<td>1</td>
<td>29</td>
<td>25.2</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2</td>
<td>86</td>
<td>74.8</td>
</tr>
<tr>
<td>Chronic Leg Ulcer Lifetime Incidence</td>
<td>Yes</td>
<td>1</td>
<td>2</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2</td>
<td>113</td>
<td>98.3</td>
</tr>
<tr>
<td>Priapism Lifetime Incidence</td>
<td>Yes</td>
<td>1</td>
<td>2</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>2</td>
<td>113</td>
<td>98.3</td>
</tr>
</tbody>
</table>

As shown in Table 4.2, it was observed that majority of the SCD patients assessed did not have any of the associated lifetime incidence related to CVD with a proportion of at least 95%, acute chest syndrome with a proportion of at least 80%, pneumococcal meningitis with a proportion of at least 98%, gall stone with a proportion of at least 97%, osteomyelitis with a proportion of at least 74%, chronic leg ulcer with a proportion of at least 98% and priapism with a proportion of at least 98%. The results of the distribution of the numeric clinical data is also presented in Table 3 in terms of the minimum, maximum, mean and standard deviation of the features assessed.
### Table 3: Description of the Numeric Attributes collected

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present Age (in years)</td>
<td>1</td>
<td>15</td>
<td>6.60</td>
<td>3.795</td>
</tr>
<tr>
<td>Age at first diagnosis (months)</td>
<td>4</td>
<td>156</td>
<td>29.62</td>
<td>28.113</td>
</tr>
<tr>
<td>Frequency of Painful Crisis</td>
<td>0</td>
<td>20</td>
<td>3.87</td>
<td>3.671</td>
</tr>
<tr>
<td>Frequency of Blood Transfusions</td>
<td>0</td>
<td>8</td>
<td>1.24</td>
<td>1.490</td>
</tr>
<tr>
<td>Frequency of Hospitalization</td>
<td>0</td>
<td>12</td>
<td>2.23</td>
<td>2.271</td>
</tr>
<tr>
<td>Spleen size (in cm)</td>
<td>0</td>
<td>18</td>
<td>4.29</td>
<td>4.432</td>
</tr>
<tr>
<td>Liver size (in cm)</td>
<td>0</td>
<td>11</td>
<td>3.96</td>
<td>2.921</td>
</tr>
<tr>
<td>Hematocrit Level (PCV) (%)</td>
<td>6.0</td>
<td>37.0</td>
<td>23.08</td>
<td>4.620</td>
</tr>
<tr>
<td>White Blood Cell (WBC) Count (/mm$^3$)</td>
<td>55</td>
<td>87000</td>
<td>15111.87</td>
<td>12528.441</td>
</tr>
<tr>
<td>HbF Level</td>
<td>1.1</td>
<td>10.3</td>
<td>5.52</td>
<td>2.461</td>
</tr>
</tbody>
</table>

The results of the data collection showed that the minimum age of SCD patients assessed is 1 year with a maximum of 15 years which yielded an average age of 6 years for SCD patients assessed. It was also observed in the results that the minimum age of diagnosis of SCD was 4 months with a maximum of 156 months which yielded an average age of 29 months. The frequency of blood transfusions, hospitalization painful crisis was evaluated based on the number of episodes in the last year. The results showed that regarding the number of painful crisis in the last year, the maximum recorded was 20 painful episodes within a distribution with a mean of 3 episodes and standard deviation of 3 episodes. The results showed that regarding the number of blood transfusions in the last year, the maximum recorded was 8 transfusion episodes within a distribution with a mean of 1 episode and standard deviation of 1 episode. The results showed that regarding the number of hospitalizations episodes in the last year, the maximum recorded was 12 within a distribution with a mean of 2 episodes and standard deviation of 2 episodes.

### 4.2 Results of ensemble model formulation and simulation

This section presents the results of the process of formulating and simulating the classification models required for assessing the severity of anemia among SCD patients using 3 supervised machine learning algorithms. In one part, the classification model was formulated using the supervised machine learning algorithms in isolation. In the other part, the classification model was formulated using an ensemble of the supervised machine learning algorithms. The simulation of the classification models was simulated using a 10-fold cross validation technique via the Waikato Environment for Knowledge Analysis (WEKA).

#### Results of the formulation and simulation of isolated models

Based on the results of the application of the C4.5 Decision Trees algorithm alone for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 57 were correctly classified while 3 and 4 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 7 were correctly classified while 5 were misclassified as moderate severe cases. The
presentation of the number of correct and incorrect classification of each target class for the severity of anemia is presented in Figure 4 (left). The results of the performance of the application of the C4.5 decision trees algorithm showed an accuracy of 84.3%.

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>M</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>33</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>35</td>
<td>57</td>
<td>4</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>M</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>31</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>4</td>
<td>57</td>
<td>3</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>M</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>33</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>2</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

**Figure 4: Results of the Isolated Classifiers**

Regarding the results of the application of the naïve Bayes’ algorithm alone for model formulation it was observed that out of the 39 actual low severe cases, 31 were correctly classified while 8 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 57 were correctly classified while 4 and 3 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 7 were correctly classified while 5 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia showed an accuracy of 82.6%.

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>M</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>33</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>3</td>
<td>57</td>
<td>4</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

Regarding the results of the application of the support vector machine (SVM) using the Sequential Minimal Optimization (SMO) algorithm alone for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 60 were correctly classified while 2 were misclassified as each of low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 6 were correctly classified while 6 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia showed an accuracy of 86.1%.

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>M</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>33</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>2</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

### Results of the formulation and simulation of ensemble of 2 models

Based on the results of the application of ensemble of algorithms, using the ensemble of C4.5 Decision Trees (DT) and naïve Bayes’ (NB) algorithm for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 59 were correctly classified while 2 and 3 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 8 were correctly classified while 4 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia using an ensemble of DT and NB is presented in Figure 5 (left). The results of the performance of the ensemble of DT and NB algorithms showed an accuracy of 87%.

Regarding the results of the application of ensemble of algorithms, using the ensemble of C4.5 Decision Trees (DT) and support vector machines (SVM) algorithm for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 57 were correctly classified while 3 and 4 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 7 were correctly classified while 5 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia using an ensemble of DT and SVM is presented in Figure 5 (right). The results of the performance of the ensemble of DT and SVM showed an accuracy of 87%. 
of correct and incorrect classification of each target class for the severity of anemia using an ensemble of DT and SVM is presented in Figure 5 (center). The results of the performance of the ensemble of DT and SVM algorithm showed an accuracy of 84%.

Regarding the results of the application of ensemble of algorithms, using the ensemble of naïve Bayes’ (NB) and support vector machines (SVM) algorithm for model formulation it was observed that out of the 39 actual low severe cases, 33 were correctly classified while 6 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 56 were correctly classified while 5 and 3 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 5 were correctly classified while 7 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia using an ensemble of NB and SVM is presented in Figure 5 (right). The results of the performance of the ensemble of DT and SVM algorithm showed an accuracy of 83%.

\[
\begin{array}{ccc|c|c|c}
L & M & H & L & M & H & L & M & H \\
33 & 2 & 0 & 33 & 59 & 3 & 33 & 6 & 0 \\
6 & 3 & 0 & 57 & 4 & 7 & 6 & 0 & 0 \\
0 & 4 & 8 & 0 & 5 & 7 & 0 & 5 & 7 \\
\end{array}
\]

Figure 5: Results of the Ensemble of Two (2) Classifiers

The results of the performance of the ensemble of all the machine learning algorithms showed that out of the 39 actual low severe cases, 32 were correctly classified while 7 were misclassified as moderate severe cases. Out of the 64 actual moderately severe cases, it was observed that 59 were correctly classified while 3 and 2 were misclassified as low and high severe cases respectively. Out of the 12 actual high cases, it was observed that 7 were correctly classified while 5 were misclassified as moderate severe cases. The presentation of the number of correct and incorrect classification of each target class for the severity of anemia using an ensemble of DT, SVM and NB is presented in Figure 6. The results of the performance of the ensemble of DT, SVM and NB algorithms showed an accuracy of 85%.

\[
\begin{array}{ccc|c|c|c}
L & M & H & L & M & H & L & M & H \\
32 & 3 & 0 & 33 & 60 & 0 & 33 & 0 & 0 \\
7 & 5 & 2 & 56 & 3 & 6 & 5 & 7 & 0 \\
0 & 5 & 7 & 0 & 5 & 7 & 0 & 5 & 7 \\
\end{array}
\]

Figure 6: Results of the Ensemble of the Three (3) Classifiers

4.3 Results of ensemble model validation

Based on the results of the performance of the DT algorithm, it was observed that the model developed classified correctly on average about 84% of the actual cases alongside misclassification rate of an average of 14% of actual cases. The results of the model developed using the DT also showed that an average of 84% of the
predictions made by the algorithm was correct. Based on the results of the performance of the NB algorithm, it was observed that the model developed classified correctly on average about 83% of the actual cases alongside misclassification rate of an average of 16% of actual cases. The results of the model developed using the NB also showed that an average of 82% of the predictions made by the algorithm was correct.

Based on the results of the performance of the SVM algorithm, it was observed that the model developed classified correctly on average about 86% of the actual cases alongside misclassification rate of an average of 14% of actual cases. The results of the model developed using the SVM also showed that an average of 86% of the predictions made by the algorithm was correct. It was also observed from the results of the isolated algorithms that among the adopted algorithms for the classification of the severity of SCD patients, SVM had the highest capability of predicting correctly the severity of anemia and the lowest ability of misclassifying the severity of SCD patients.

Based on the results of the performance of the ensemble of DT and NB algorithms, it was observed that the model developed classified correctly on average about 87% of the actual cases alongside misclassification rate of an average of 12% of actual cases. The results of the model developed using the ensemble of DT and NB also showed that an average of 87% of the predictions made by the algorithm was correct.

Based on the results of the performance of the ensemble of DT and SVM algorithms, it was observed that the model developed classified correctly on average about 84% of the actual cases alongside misclassification rate of an average of 14% of actual cases. The results of the ensemble model developed using the DT and SVM also showed that an average of 84% of the predictions made by the algorithm was correct.

Based on the results of the performance of the ensemble of NB and SVM algorithms, it was observed that the model developed classified correctly on average about 84% of the actual cases alongside misclassification rate of an average of 15% of actual cases. The results of the model developed using the ensemble of NB and SVM also showed that an average of 83% of the predictions made by the algorithm was correct.

It was also observed from the results of the ensemble of 2 algorithms that among the adopted algorithms for the classification of the severity of anemia among SCD patients, using and ensemble of DT and SVM had the highest capability of predicting correctly the severity of anemia and the lowest ability of misclassifying the severity of SCD patients.

Also, based on the ensemble of 3 algorithms using DT, SVM and NB, it was observed that the model developed classified correctly on average about 85% of the actual cases alongside misclassification rate of an average of 15% of actual cases. The results of the model developed using the ensemble of DT, SVM and NB also showed that an average of 85% of the predictions made by the algorithm was correct. The results of the study showed that out of all the models developed for the classification of the severity of anemia among SCD patients that the adoption of the ensemble of DT and NB showed the best performance out of all the proposed combinations. The results showed that although in isolation, NB and DT did not do as well as did SVM but what could not be achieved in isolation was compensated for in combination.

On the other hand, the ensembles that were developed using the combination of SVM with either NB or DT did not produce results better than that produced by the SVM in isolation. It was also observed from the results that ensemble of DT with SVM did not shows any improvement in performance over the use of DT in isolation while the ensemble of the 3 algorithms was observed to be better than the use of NB and DT in isolation. It was also observed for the results that using the ensemble of the 3 algorithms had better performance compared to the use of the ensemble of DT and SVM and that of NB and SVM. In all, the ensemble of DT and NB was observed to produce the best results among the various combination of ensemble models for the 3 supervised machine learning algorithms adopted in this study due to its high TP rate and Precision and low FP rates. The summary of the model validation results is presented in Table 4.
Table 4.4: Results of the Evaluation of the Performance of Classification Models

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Correct</th>
<th>TP rate</th>
<th>FP rate</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5 Decision Trees (DT)</td>
<td>84.348</td>
<td>97</td>
<td>0.843</td>
<td>0.137</td>
<td>0.844</td>
</tr>
<tr>
<td>Naïve Bayes (NB)</td>
<td>82.609</td>
<td>95</td>
<td>0.826</td>
<td>0.163</td>
<td>0.827</td>
</tr>
<tr>
<td>Support Vector Machines (SVM)</td>
<td>86.087</td>
<td>99</td>
<td>0.861</td>
<td>0.142</td>
<td>0.862</td>
</tr>
<tr>
<td>Ensemble I (DT + NB)</td>
<td>86.957</td>
<td>100</td>
<td>0.870</td>
<td>0.121</td>
<td>0.872</td>
</tr>
<tr>
<td>Ensemble II (DT + SVM)</td>
<td>84.348</td>
<td>97</td>
<td>0.843</td>
<td>0.137</td>
<td>0.844</td>
</tr>
<tr>
<td>Ensemble III (NB + SVM)</td>
<td>83.478</td>
<td>96</td>
<td>0.835</td>
<td>0.145</td>
<td>0.833</td>
</tr>
<tr>
<td>Ensemble IV (DT + NB + SVM)</td>
<td>85.217</td>
<td>98</td>
<td>0.852</td>
<td>0.146</td>
<td>0.854</td>
</tr>
</tbody>
</table>

5. Conclusion

The study concluded that a number of demographic and clinical variables were associated with the risk of anaemia based on the severity of SCD among patients. The study concluded that the data collected contained a majority of Moderate risk cases followed by Low risk and High risk cases. The study concluded that following the process of model formulation and simulation using the WEKA simulation environment that by using a combination of classifiers, it was observed that a better performance was detected compared to using the classifiers in isolation for model development.

The study concluded that the process of model classification using the isolated classification algorithms showed that among the isolated model, the application of SVM yielded the best results among the 3 selected algorithms selected for this study. On the other hand, following the performance of the SVM among the isolated algorithms is the application of the C4.5 decision trees algorithm followed by naïve Bayes’ algorithm.

The study concluded that the model development using an ensemble of two (2) algorithms revealed that the performance of the ensemble which applied decision trees outperformed the performance of the model formulated using the isolated decision trees algorithms. The results however showed that both ensemble models created using SVM with NB and DT did not perform well as did the performance of the model formulated using the isolated SVM algorithm. Also, the results showed that all the ensemble of 2 algorithms outperformed the use of naïve Bayes’ classifier in isolation.

The study concluded that using an ensemble of 3 classifiers, a performance better than that using either DT or NB in isolation and that of using and ensemble of either DT+SVM or NB+SVM was determined. On a general note, the results showed that the best classification model for determining the severity of anemia among SCD patients was developed using an ensemble of DT and NB algorithms. The study concluded that the predictive model for the risk of anaemia with the best performance was the ensemble model which combined the C4.5 decision trees (DT) and the naïve Bayes’ (NB) classifiers.

References


A General Study on Langevin Equations of Arbitrary Order

S. Harikrishnan1,*, E. M. Elsayed2, K. Kanagarajan1

1Department of Mathematics, Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Coimbatore-641020, India.

2Department of Mathematics, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia.

2Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt.

Abstract

In this paper, the broad study depends on Langevin differential equations (LDE) of arbitrary order. The fractional order is in terms of ψ-Hilfer fractional operator. This work reveals the dynamical behaviour such as existence, uniqueness and stability solutions for LDE involving ψ-Hilfer fractional derivative (HFD). Thus the fractional LDE with boundary condition, impulsive effect and nonlocal conditions are taken in account to prove the results.

Keywords: Langevin differential equations, Fractional calculus, Existence, Stability

1. Introduction

The physical phenomena in fluctuating environments are described using Langevin equation. The Langevin equation is a powerful tool for the study of dynamical properties of many interesting systems in physics, chemistry and engineering. The generalized Langevin equation was introduced by Kubo in 1966. Since then the generalized LDE has become a searing research topic. The literature on LDE is huge; dynamical analysis and important results can be seen in [1, 2, 3, 7, 9]. The fractional derivatives make the fractional-order models more realistic than the classical integer-order model. Since the development of fractional calculus in various fields such as thermodynamics, biophysics, aerodynamics, viscoelasticity, capacitor theory, etc., there was been an intensive development in fractional derivative with singular and non-singular kernels. The Riemann-Liouville, Caputo, Hadamard, Hilfer, etc., are just a few fractional derivatives. Later HFD is fused with ψ-fractional derivative with kernel of function and pulled out a new fractional derivative known as ψ-HFD. The ψ-HFD integrate numerous fractional derivative with their properties are discussed in [10]. The dynamical behaviour and development of differential equation with different fractional derivatives, we refer to [5, 6, 8, 11, 12, 13]. Motivated by the works mentioned, here LDE with ψ-HFD involving boundary, impulsive and nonlocal conditions are studied. Stability criteria is an important aspect of differential equations of arbitrary order. The stable solution of fractional LDE is provided by utilizing the idea provided by Ulam. Thus, we discuss the generalized Ulam-Hyers-Rassias (g-UHR) stable for fractional LDE.