MISSING DATA IMPUTATION FRAMEWORK FOR EARLY CHILDHOOD LONGITUDINAL DATA: A STUDY CASE ON NCDRC DATA

ABDULLAH HUSSEIN ABDULLAH AL-AMOODI

UNIVERSITI PENDIDIKAN SULTAN IDRIS

2019

MISSING DATA IMPUTATION FRAMEWORK FOR EARLY CHILDHOOD LONGITUDINAL DATA: A STUDY CASE ON NCDRC DATA

ABDULLAH HUSSEIN ABDULLAH AL-AMOODI

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF ART, COMPUTING AND CREATIVE INDUSTRIES
SULTAN IDRIS EDUCATION UNIVERSITY

2019

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

UNIVERSITI PENDIDIKAN SULTAN IDRIS
اونيۏرسيتي ڤنديديقن سلطان ادريس
SULTAN IDRIS EDUCATION UNIVERSITY

UPSI/IPS-3/BO 32
Pind: 00 m/s: 1/1

Please tick (✓)

Project Paper	
Masters by Research	
Masters by Mix Mode	
Ph.D.	✓

INSTITUTE OF GRADUATE STUDIES

DECLARATION OF ORIGINAL WORK

This declaration is made on the 12/11/2019

i- Student's Declaration:

I Abdullah Hussein Abdullah Al-Amoodi-P20171000971-Faculty of Art, Computing, and Creative Industry hereby declares that the dissertation /thesis for Doctor of Philosophy titled "Missing Data Imputation Framework for Early Childhood Longitudinal Data: A Study Case on NCDRC Data" is my original work. I have not plagiarized from any other scholar's work and any sources that contain copyright had been cited properly for the permitted meanings. Any quotations, excerpt, reference or re-publication from or any works that have copyright had been clearly and well cited.

____

Signature of the student

ii- Supervisor's Declaration:

I Dr. Bilal Bahaa Zaidan hereby certify that the work entitled, "Missing Data Imputation Framework for Early Childhood Longitudinal Data: A Study Case on NCDRC Data" was prepared by the above-named student, and was submitted to the Institute of Graduate Studies as a partial / full fulfillment for the conferment of the requirements for Doctor of Philosophy (By Research), and the aforementioned work, to the best of my knowledge, is the said student's work.

____________________ ________________________

Date Signature of the Supervisor

UNIVERSITI PENDIDIKAN SULTAN IDRIS

SULTAN IDRIS EDUCATION UNIVERSITY

UPSI/IPS-3/BO 31
Pind.: 01 m/s:1/1

INSTITUT PENGAJIAN SISWAZAH /

INSTITUTE OF GRADUATE STUDIES

BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM

Tajuk / *Title*: **Missing Data Imputation Framework for Early Childhood Longitudinal Data: A Study Case on NCDRC Data**

No. Matrik /*Matric No.*: **P20171000971**

Saya / *I* : **Abdullah Hussein Abdullah Al-Amoodi**

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana)* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-
acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.
   The thesis is the property of Universiti Pendidikan Sultan Idris

2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.
   Tuanku Bainun Library has the right to make copies for the purpose of reference and research.

3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.
   The Library has the right to make copies of the thesis for academic exchange.

4. Sila tandakan (√) bagi pilihan kategori di bawah / Please tick (√) from the categories below:-

☐	**SULIT/*CONFIDENTIAL***	Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. / *Contains confidential information under the Official Secret Act 1972*
☐	**TERHAD/*RESTRICTED***	Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. / *Contains restricted information as specified by the organization where research was done.*
☐	**TIDAK TERHAD / *OPEN ACCESS***	

____

(Tandatangan Pelajar/ Signature)

Tarikh: ________________

(Tandatangan Penyelia / *Signature of Supervisor*)

& (Nama & Cop Rasmi / *Name & Official Stamp*)

Catatan: Jika Tesis/Disertasi ini **SULIT @ TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

Notes: If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the related authority/organization mentioning the period of confidentiality and reasons for the said confidentiality or restriction.

ACKNOWLEDGEMENT

"بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ"

"*In the name of Allah the most gracious the most merciful*"

The past 2 years have been thus far the most challenging, interesting, and rewarding part of my life. I am very thankful and grateful that I have crossed paths with many wonderful people who have helped me in many ways in my pursuit of completing my PhD thesis at Sultan Idris Educational University (UPSI) in Malaysia.

Above all, I want to thank Allah for all his grace and mercy that makes it possible for me to produce this work. Without Allah's mercy and compassion, none of this would be possible including my very own existence. In addition, many individuals have been by my side and their stand should be acknowledged.

First and foremost, I would like express my sincere gratitude for two individuals; my supervisor, my best friend, and my role model; Dr Bilal Bahaa Zaidan and my academic advisor; Dr Aws Alaa Zaidan. They have been making my life extremely challenging and hellish since day one. Nevertheless, they were able to form a version of me I never thought existed. Their assistance and dedicated involvement in every step have not left a free moment for me to waste. Even if I assign this entire acknowledgement just for Dr Bilal alone, it will not be a fraction of my admiration, my gratitude and my respect for him. His countless 5AM late nights with me on my research have been stuck in my mind for the rest of my life. The following picture speaks for itself;

Thanks Dr Bilal for being integrated part of life, Thanks for all the nights we hang out together, thanks for you and your wife Roqaiah for all the delicious meals she prepared for us, thanks for all our ups and downs and thanks for your brotherhood. I could not have imagined having a better supervisor and mentor, let alone true bother in my PhD.

I would also like to thank my family specially my two uncles: Mohammed, Ali and my father Hussein, and my cousin Turki for encouraging me to pursue my PhD. In addition, behind every great man, a great woman and I was blessed with many including my mother Noor, My Auntie Safiah and a great woman who I consider as a mother and has huge part in my life Nasrah. For all of you, I love you so much and thanks for your support.

I must express my very profound gratitude to my dearest friends; Ghailan, for being more than an older brother, and for standing my jokes and silly comments about him. Also, my friend Chyad for being more than a brother, for making me feel like family, and for bearing with me through thick and thin. To you my brother, a debt I could never repay, also my brother Salem who started the journey with me since our first days before our undergraduate studies, for his brotherhood, I could never ask for more. For my Friend Azzam, for introducing me to both Dr Bilal and Dr Aws, thanks brother, your act of kindness has changed my life.

I would also like to acknowledge my co-supervisor; Dr Suzani Samuri for helping me out with the NCDRC data. I cannot forget both my brothers Osama and Mohammed Asim for their assistance in my data selection. Also Dr Amilea from the International Islamic University Malaysia for her assistance during my statistical analysis. I am gratefully indebted to her for the very valuable comments on my work till its finalization.

I thank my fellow lab mates; Khaled, Moceheb, Ahmed Albahri, Mohammed Talal, Ahmed Marwan, Fawaz, Fayez, Omar, Mohammed Aktham, Mahmoud, Ali *Khashab* , Mussab, and Abo Qays for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last two years.

A very special gratitude goes out to all my friends at “*No Pain No Gain*” WhatsApp group members for their encouragement.

Finally, I also would like to thank all my brothers and sisters for their support throughout the years. Haters and critics are also worth thanking for their harsh words and inappropriate remarks which fueled my will to never give up and to never be arrogant in my academic life.

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

DEDICATION

To Dr Bilal Bahaa Zaidan

For believing in me and for being with me since the beginning.

To my Future wife whom I am yet to meet

I hope by the time you read this, you become damn proud.

To all my friends in Saudi Arabia and in Malaysia

إلى اليد الطاهرة التي أزالت من أمامنا أشواك الطريق ورسمت المستقبل بخطوط من الأمل والثقة
إلى الذي لا تفيه الكلمات والشكر والعرفان بالجميل
والدي الحبيب: حسين بن عبدالله العمودي
إلى من ركع العطاء أمام قدميها
وأعطتنا من دمها وروحها وعمرها حبا وتصميما ودفعا لغدٍ أجمل
إلى الغالية التي لا نرى الأمل إلا من عينيها
أمي الحبيبة: نور عبدالرحمن العمودي
إلى من كانتا سندي وعوني في ايامي وغربتي
إلى من كانتا دوما بقربي رغم بعد المسافات
إلى الغاليتين
خالتي العزيزة: صفية عبدالرحمن العمودي
وأمي الغالية ناصرة بنت علي
إلى اخواني واخواتي الغالين
أحمد , عبدالرحمن , إيمان وأمل
أهديكم جميعا هذا النجاح

ABSTRACT

This research aims to develop an imputation framework for the National Childhood Development Research Centre (NCDRC)'s missing data. Missing data and other associated issues, such as outliers, time points, noise, and continuity, were the main challenges in this research. The nature of the NCDRC dataset was not consistent with those reported in the literature, with the latter being more randomly scattered and copious and having no patterns, making it difficult to find and select relevant experimental data. The VIseKriterijumska Optimizacija Kompromisno Resenje (VIKOR) method was utilized to select the best continuous portion of Body Mass Index (BMI) data over 182 different portions, which accounted for 911 participants (i.e. children with complete records) over seven (7) continuous time points. Three different machine learning algorithms to impute the missing data were tested and evaluated, namely K-nearest Neighbour (KNN), Naïve Bayes (NB), and Decision Tree (DT). Three evaluation performance indicators, namely t-test, Coefficient of Determination, and Root Mean Square Error, were used in the experiment using three configurations based on 5%, 10%, and 15% missing data. The results of the experiment showed that KNN's performance scores were significantly higher than those of the other algorithms. Out of all scores, KNN achieved 95.23% of the scores, followed by NB with 94.04% and DT with 83.33 %, clearly indicating that KNN outperformed DT and NB in the imputation of missing data. In conclusion, the main finding suggests that the KNN algorithm is the most effective algorithm for imputing missing data. The implication of this study is that practitioners, especially NCDRC's personnel, can use the proposed missing data imputation framework to help impute missing data of similar datasets.

RANGKA KERJA IMPUTASI DATA HILANG UNTUK DATA LONGITUD AWAL KANAK-KANAK: SATU KAJIAN KES TERHADAP DATA NCDRC

ABSTRAK

Kajian ini bertujuan untuk membangunkan satu rangka kerja imputasi untuk Pusat Kajian Pembangunan Awal Kanak-kanak Kebangsaan (NCDRC). Data hilang dan isu-isu yang berkaitan, seperti *outlier*, titik masa, kebisingan, dan kesinambungan, merupakan cabaran yang dihadapi dalam kajian ini. Ciri set data NCDRC adalah tidak konsisten dengan set data yang dilaporkan dalam literatur, di mana ianya lebih bertaburan secara rawak dan bersaiz besar dan tidak mempunyai corak yang jelas bagi memudahkan pencarian dan pemilihan data eksperimen. Kaedah *VIseKriterijumska Optimizacija Kompromisno Resenje* (VIKOR) digunakan untuk memilih bahagian selanjar yang terbaik untuk data Indeks Jisim Badan (BMI) yang merangkumi 182 bahagian yang berlainan melibatkan 911 peserta (kanak-kanak dengan rekod yang lengkap) meliputi tujuh (7) titik-titik masa selanjar. Tiga algoritma pembelajaran mesin untuk imputasi data hilang diuji dan dinilai, iaitu *K-nearest Neighbour* (KNN), *Naïve Bayes* (NB), dan *Decision Tree* (DT). Tiga penunjuk prestasi penilaian, iaitu *t-test*, *Coefficient of Determination*, dan *Root Mean Square Error*, digunakan dalam eksperimen berdasarkan beberapa konfigurasi yang melibatkan kehilangan data sebanyak 5%, 10%, dan 15%. Dapatan menunjukkan skor prestasi KNN adalah lebih tinggi dari skor algoritma yang lain. Daripada semua skor yang terlibat, KNN memperoleh 95.23% daripada skor berkenaan, diikuti dengan NB dengan 94.04% dan DT dengan 83.33%. Ini menunjukkan KNN berprestasi lebih baik lagi berbanding DT dan NB dalam imputasi data hilang. Kesimpulannya, dapatan menunjukkan algoritma KNN adalah merupakan algoritma yang terbaik bagi imputasi data hilang. Implikasi kajian ini membolehkan para pengamal, terutamanya kakitangan NCDRC, menggunakan rangka kerja imputasi data hilang yang dibangunkan untuk menggantikan data yang hilang dalam sesuatu set data.

TABLE OF CONTENT

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

LIST OF TABLES

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

ACE	Adverse Childhood Experiences
AHP	Analytic Hierarchy Process
AI	Artificial Intelligence
ANP	Analytic Network Process
ASD	Autism Spectrum Disorder
BMI	Body Mass Index
BWM	Best-Worst-Method
CF	Cystic Fibrosis
CFA	Confirmatory Factor Analyses
DM	Decision Matrix
DNA	Deoxyribonucleic Acid
DT	Decision Tree
ECCE	Childhood Care And Education
EEG	Electroencephalography
EMG	Electromyography
FIML	Full Information Maximum Likelihood
fNIRS	Functional Near-Infrared Spectroscopy
GEE	Generalized Estimating Equation
KNN	K-Nearest Neighbor
LPA	Latent Profile Analysis
MCDM	Multi Criteria Decision Making
MEW	Multiplicative Exponential Weighting

MI	Multiple Imputation
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NB	Naïve Bayes
NCDRC	National Child Development Research Center
PD	Professional Development
SAW	Simple Additive Weighting
SES	Socioeconomic Status
SLR	Systematic Literature Review
SVM	Support Vector Machine
TOPSIS	The Technique For Order Preference By Similarity To Ideal Solution
VIKOR	Vlse Kriterijumska Optimizacija Kompromisno Resenje
WoS	Web Of Science
WPM	Weighted Product Method
WSM	Weighted Sum Model

LIST OF APPENDICES

CHAPTER 1

RESEARCH BACKGROUND

1.1 Introduction

This chapter explains the research background aspect of this thesis, it is meant to highlight different areas and points which contribute to the understanding of this thesis's topic. Among the points covered in this chapter are the research background which informed the reader about the origin of the topic, followed by the state of problem which discusses how the problem in this dissertation emerges. Furthermore, other important highlights are addressed including research objectives, research questions and research scope. As for the last details which summarize this thesis layout.

1.2 Research Background

Early childhood is the period when most of children transitions take place; this period is a significant influencer of child development as children progress into adolescence (Caemmerer & Keith, 2015) and adulthood (Kim, Choi, & Kim, 2014). Early childhood is an intriguing research in the academic world that warrants considerable attention (Girard, Pingault, Doyle, Falissard, & Tremblay, 2017). Early childhood was investigated in many domains, such as social and medical domains. Researchers emphasized the importance of this period in shaping many aspects of children's lives, especially brain development as discussed in (Keyser, Ahn, & Unick, 2017) and (M. Wang & Saudino, 2013). This period plays a significant role in shaping other aspects of childhood development, such as growth (Matos et al., 2017; Schott, Crookston, Lundeen, Stein, & Behrman, 2013), emotions (Keyser et al., 2017; Kim et al., 2014; Long, Benischek, Dewey, & Lebel, 2017; M. Wang & Saudino, 2013), socialization and behavior (Girard et al., 2017; Hardee et al., 2013; Long et al., 2017; Matos et al., 2017; Taveras, Rifas-Shiman, Bub, Gillman, & Oken, 2017) and health (Matos et al., 2017). Apart from children's development aspect, this period plays a great role on child's skills including cognitive (Kim et al., 2014; Long et al., 2017; Taveras et al., 2017), perception (Y. Zhang et al., 2017), inhibitory control (Gagne & Saudino, 2016), executive function (Meuwissen & Englund, 2016), language (Girard et al., 2017) and education (B. Jensen, Jensen, & Rasmussen, 2017; X. Zhang & Lin, 2015). Early childhood also represents great risk, wherein many neurodevelopmental disorders emerge (Long et al., 2017) in addition to the internalization and externalization of problems in children (F. Li & Godinet, 2014). Family bonds between parents and children are formed during this time (De Luca, Yueqi, & Padilla, 2017). Literature shows that early childhood research is gaining significant interest, which is recognized as a hot topic (Conroy & Harcourt, 2009). Several researchers conducted data analysis on this domain and obtained good results (e.g. identifying special patters in a particular skills). Towards this end, data analysis is essential in the study of early childhood in order to understand this period in its best shape.

Data analysis plays a significant role in science research. Researchers worldwide utilized different data analysis measures in their respected fields for different reasons. Researchers tend to work with different analysis methods for various scientific purposes, such as evaluation (Aubert-Broche et al., 2013), investigations (Erdoğan, Ener, & Arıca, 2013; Y. Zhang et al., 2017), data modelling (Greenwood et al., 2013) and prediction (Staff, Maggs, Cundiff, & Evans-Polce, 2016). Others highlighted the role of data analysis for its contribution to research (Meuwissen & Englund, 2016), answering research questions (Speirs et al., 2016) and the relationship of findings (Mills-Koonce et al., 2015). While others developed data analysis for social purposes, particularly, understanding children's related norms, such as health (Schleider, Abel, & Weisz, 2015), tracking and describing changes (Guevara, van Dijk, & van Geert, 2016; R. Miller, 2017); reducing biases (Field, 2017; C.-W. Liu et al., 2017; Paternina-Caicedo et al., 2015) and their contribution to lack of consensus (Buckley et al., 2015). Due to the variety of data with respect to its nature, different analytical approaches are required for different types of data. Accordingly, the nature and type of data used for the analysis play a major role in any data analysis approach.

Data type is also a significant part of any study, which is equally important as data analysis because they complete each other. Data type alone does not provide sufficient information without proper analysis. The same goes for data analysis, which would not provide information without proper data. It has been noticed that data types in early childhood studies represent an occurrence or a situation for children at that time period. Data type may not be a significant indicator of how this period is observed given the fact that data related to children are not recorded by them (i.e. children), but they are recorded by parents or other caregivers (Netsi et al., 2017; Strobino et al., 2016; Trautmann, Alhusen, & Gross, 2015). A longitudinal type of data emerged in literature, which enables researchers to observe children for many years. This type of data aids researchers across the majority of early childhood studies for various purposes, such as evaluation (Aubert-Broche et al., 2013), improvement of study models (Sadeghi et al., 2013), provision of consistent estimates (Matos et al., 2017), maximization of statistical power (Ducharme et al., 2016; Royal-Thomas, McGee, Sinha, Osmond, & Forrester, 2015) and maintenance of data records (Long et al., 2017). Longitudinal data also vary in terms of usage; some researchers used these data for answering research questions (Buckley et al., 2015), conducting investigations (Simpkin et al., 2017; Tamilia, Formica, Scaini, & Taffoni, 2016), performing comparisons (McCormick, O'Connor, Cappella, & McClowry, 2013), testing models (Shaw et al., 2014; Torres, Domitrovich, & Bierman, 2015), providing clarifications (Bellin et al., 2013), examining effects (C. E. Baker & Iruka, 2013; Carlson, Sonderegger, & Bane, 2014; Heatly, Bachman, & Votruba-Drzal, 2015; Maslow et al., 2017) and determining findings (Jeon, Peterson, & DeCoster, 2013). Longitudinal data therefore is a very capable and rich type of data filled with large volume of information, such data are not like the usual type, since it requires commitment and large financial support. This data type is not an easy one to start and maintain, since it requires huge resources which can be found in governments and large scale institutions (i.e. UK Millennium Study). Despite its large and vast academic and scientific worth, such type of data might lost its significance (brightness) and not meet its expectations due to human and technical errors (Van den Broeck, Cunningham, Eeckels, & Herbst, 2005) (e.g. missing data, outliers, unprofessional reporting, etc.).

Missing data is an inevitable occurrence associated with the data collection process, especially when the data collected are huge and contains large number of inputs. This issue can cause several drawbacks affecting the findings later on. Among the drawbacks of the missing data comes the possibility of bias findings (R. Miller, 2017; Vandecandelaere, Vansteelandt, De Fraine, & Van Damme, 2016), reducing the sample size (Singh, Winsper, Wolke, & Bryson, 2014), excluding data (Gordon, Colaner, Usdansky, & Melgar, 2013) and the inability to understand changes in the data (Derrington et al., 2013). However, missing data should be taken into account specially when dealing with repeated measurement (L.-W. Chen et al., 2016). The importance of dealing with the missing data should begin during the data collection stage, and all suitable environments should be setup in advance to encourage participants to fill up the data efficiently and reduce the ratio of missing occurrences. Missing data thus can be handled by means of statistical procedures, by means of machine learning, or by elimination.

1.3 Problem Statement

Observing the academic literature suggested different challenges when it comes to early childhood research. Most of these challenges are (in away or other) related to the data as seen in Figure 1.1. Five classes of challenges are summarized, namely, processing of data, collection/acquisition of data, nature of data, the procedure of reporting, and finally sample size. Each of these classes are associated with a number of aspects. The first challenge reported in the academic literature is concerned with elements of data collection, includes missing data (i.e. refers to incomplete data received by participants) (Lê, Roux, & Morgenstern, 2013; Pluymen et al., 2016; Singh et al., 2014), bias reporting of data, including reporting data of children by their parents (C. E. Baker & Iruka, 2013; De Luca et al., 2017; Green, Tarte, Harrison, Nygren, & Sanders, 2014; Hartman et al., 2016; Jones, Champion, & Woodward, 2013; Keyser et al., 2017; McKelvey, Selig, & Whiteside-Mansell, 2017; Meuwissen & Englund, 2016; Nath, Russell, Kuyken, Psychogiou, & Ford, 2016), their teachers (Brotman et al., 2016; Caemmerer & Keith, 2015; Heatly et al., 2015), children themselves (Aschengrau et al., 2016; Contzen, Meili, & Mosler, 2015; Sunny et al., 2017), and their parents (C. E. Baker & Iruka, 2013; Meuwissen & Englund, 2016; Nath et al., 2016), and finally data reporting by unprofessional (i.e. refer to the commitment of personnel who are in charge of gathering the data and key it in the system correctly (Tharayil et al., 2017). Bias reporting can effect integrity of data, and reducing the benefit of analysis result. Another challenge is presented within the nature of data, includes the following aspects, structure, (i.e. refers to the structure format of the data) (Caemmerer & Keith, 2015), and accuracy (i.e. refers to certainty of reported records in their actual and scheduled time) (Sunny et al., 2017). The third challenge identified in literature involves different elements with respect to the method of data sampling. The first is sample size, (i.e. population representation, and number of participants) (Ifflaender, Rüdiger, Konstantelos, Wahls, & Burkhardt, 2013; Yuan et al., 2016; Zare, Rezvani, & Benasich, 2016) , another aspect associated with data sampling involves the time points (i.e. refers to participants in survey with uncompleted/limited to one or few time points) (Kremer, Flower, Huang, & Vaughn, 2016; McCormick, O’Connor, & Barnes, 2016) . Different challenge involves different aspects linked to processing of data, includes statistical power (Aschengrau et al., 2016; McDonald et al., 2013; Tamayo, Manlhiot, Patterson, Lalani, & McCrindle, 2015), noise (i.e. refers to data entries which are not relevant to the rest of entries like 0 value where it is supposed to be a number) (S. T. Baker, Leslie, Gallistel, & Hood, 2016; Bhattacharya, Ehrenthal, & Shatkay, 2014; Tharayil et al., 2017), and finally outliers (i.e. refers to extreme data records which cannot be achieved by any means) (Agyei, van der Weel, & van der Meer, 2016; Brown, Gyllenberg, Hinkka-Yli-Salomäki, Sourander, & McKeague, 2017; Long et al., 2017).

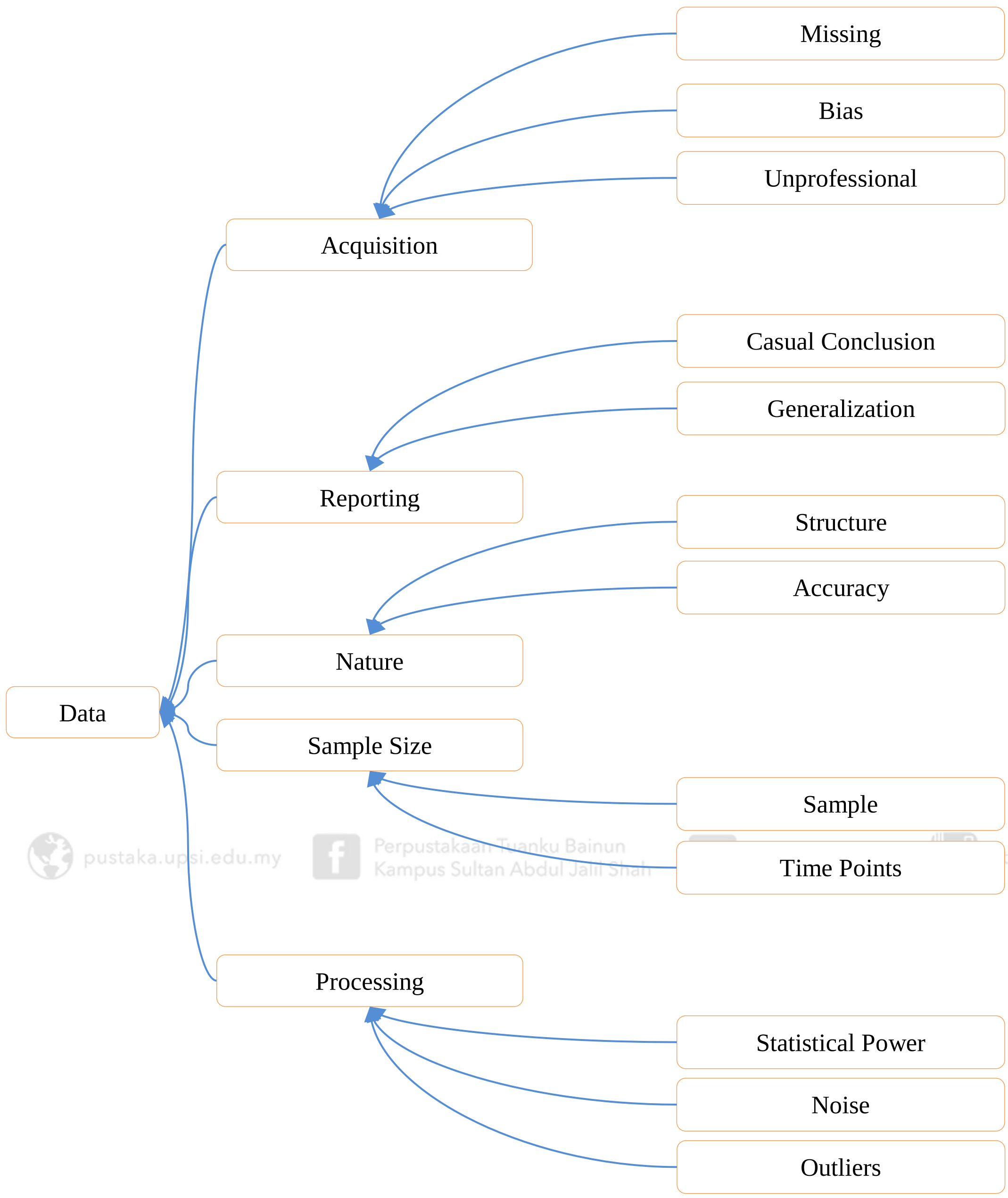

Figure 1.1. Problem Statement Main Components

The last challenge present those aspects linked with the way data is reported. This includes inability to draw casual conclusions (i.e. do not meet the expectations of the findings, or uncertain conclusions) (Heatly et al., 2015; Yang & Yang, 2015) , and the inability to generalize the findings to whole population (Guevara et al., 2016; Zare et al., 2016).

These issues are deemed significant and play an important role in the findings. A brief screening has been conducted for early childhood, data analysis and data types studies considering the scope of age between infancy and 5 years old as per the available data in National Child Development Research Center (NCDRC). It is found out that out of *233* studies identified, (*n=127/233*) was done in the in the United State of America followed by *13* studies from the United Kingdom. Within the scope of this research, Asian countries produce limited number of research articles with (*n=23/233*). It seems that the interest in Asia is considered slim compared with countries like USA, and even among the ones in Asia, Malaysia had no study found, see section 2.5.1. In Malaysia, NCDRC is a very capable center to produce strong findings associated with various early childhood topics and the overall records are estimated to be around *96000* records for more than *16000* child. However, as any source of big longitudinal data, there are some issues that hamper the integrity of such massive records. In NCDRC, the records available are around *96000* records but the ones with no missing value and completed are only around *168* records across all the time points for *12* child. Similar to academic literature, most of the issues reported in the academic literature are identified within NCDRC dataset, in particular, missing data, unprofessional data reporting, bias, outliers, noise and so on.

It is clearly identified that data holds major share of issues reported in previous literature. The less missing data, the larger sample size, the more generalizable findings, the more representative samples. Nevertheless, other benefits including, producing highly accurate and solid findings

Looking back at means of imputing missing data, it is identified that most of them follow statistical analysis approaches. Moreover, most of the previous attempts only imputed data within small scales without understanding the overall nature of data and its variables. Bearing that in mind, if dataset was understood thoroughly, a big chance that it could be utilized for its maximum (Hahn & Haisken- DeNew, 2013) and therefore rebuilt considering the correlations between its variables. Therefore, the idea of imputing data via statistical means is excluded. On the other hand, machine learning, (even though not widely presented within such application for early childhood studies at least in the selected set of papers in this systematic literature review) shows promising results in imputing data (Jerez et al., 2010). It is believed that the utilization of machine learning can aid in the missing data imputation and maintain data integrity in the process.

For several reasons mentions in section 3.8, existing missing data imputation approaches (i.e. machine learning imputation) are not directly applicable to our dataset. A generic and unique framework for imputation of missing data across different dataset with different characteristics is yet to be identified in the academic literature. Thus, three machine learning algorithms (K-nearest Neighbor, Decision Tree, and Naïve Bayes) proposed for imputing missing data are to be experimented upon, with different experiments and different settings until final conclusion is observed for best performing one in this dataset.

This research is an attempt to rebuild parts of NCDRC dataset towards increasing an applicable part of the data. This can be achieved by imputing the largest missing data portion possible using machine learning without affecting the overall integrity of the dataset.

Towards this end, NCDRC would compete with great centers of early childhood studies across the glove (i.e. *UK Millennium project*) in terms of academic studies in addition to social side which may aid children development and health. Table 1.1 presents the most common previous gaps (i.e. related to this study) with their full references.

Table 1.1
Gaps with Full References

Gap		**Full References**
Inability To Generalize The Findings		(Apouey, 2016; Bellin et al., 2013; Boxum et al., 2014; Bradley-Hewitt et al., 2016; Burgess, Audet, & Harjusola-Webb, 2013; J. Y. Choi, Jeon, & Lippard, 2018; Derrington et al., 2013; Fenstermacher & Saudino, 2016; Goelman, Zdaniuk, Boyce, Armstrong, & Essex, 2014; Green et al., 2014; Guevara et al., 2016; Ifflaender et al., 2013; B. Jensen et al., 2017; Kildare & Middlemiss, 2017; Kim et al., 2014; S. J. Lee, Altschul, & Gershoff, 2015; Mallan, Fildes, Magarey, & Daniels, 2016; Mann, Mcdermott, Pan, & Hardin, 2013; McCormick et al., 2013; Merritt & Klein, 2015; Nelson et al., 2013; Price, Higa-McMillan, Kim, & Frueh, 2013; Sisson et al., 2016; Strobino et al., 2016; Tamayo et al., 2015; Taveras et al., 2017; Torres et al., 2015; Uzark, Smith, Donohue, Yu, & Romano, 2017; Xu, Wen, Hardy, & Rissel, 2016; Yoon, 2017; Zare et al., 2016; X. Zhang & Lin, 2015; Y. Zhang et al., 2017)
Biased Reporting by Parents	For Children	(Bellin et al., 2013; J.-H. Chen, 2014; Christiana, Battista, James, & Bergman, 2017; De Luca et al., 2017; Fang et al., 2014; Gibbs & Forste, 2014; Girard et al., 2017; Hermanns, Asscher, Zijlstra, Hoffenaar, & Deković, 2013; Koulouglioti et al., 2014; Kremer et al., 2016; Lê et al., 2013; Lewis, McElroy, Harlaar, & Runyan, 2016; F. Li & Godinet, 2014; Mika et al., 2016; Nam & Chun, 2014; Netsi et al., 2017; Shoda et al., 2016; Speirs et al., 2016; Strobino et al., 2016; Taveras et al., 2017; M. Wang & Saudino, 2015; Xu et al., 2016; Yoon, 2017)
	For Parents	(C. E. Baker & Iruka, 2013; De Luca et al., 2017; Green et al., 2014; Hartman et al., 2016; Jones et al., 2013; Keyser et al., 2017; McKelvey et al., 2017; Meuwissen & Englund, 2016; Nath et al., 2016)
Missing Data		(Aschengrau et al., 2016; Derrington et al., 2013; Gordon et al., 2013; Greenwood et al., 2013; Heatly et al., 2015; Lê et al., 2013; Merritt & Klein, 2015; Pluymen et al., 2016; Singh et al., 2014; H. Wang, Tian, Wu, & Hu, 2016)

(*Continue*)

Table 1.1 (*Continued)*

Gap	**Full References**
Sample Size	(Aschengrau et al., 2016; Becker, Miao, Duncan, & McClelland, 2014; Boxum et al., 2014; Brooker & Buss, 2014; Brotman et al., 2016; Burgess et al., 2013; Christiana et al., 2017; Contzen et al., 2015; Derrington et al., 2013; Gagne & Saudino, 2016; Gauld, Keeling, Shackleton, & Sly, 2014; Geisbush, Visyak, Madabusi, Rutkove, & Darras, 2015; Gordon et al., 2013; Guevara et al., 2016; Hardee et al., 2013; Hermanns et al., 2013; Hoff, Rumiche, Burridge, Ribot, & Welsh, 2014; Ifflaender et al., 2013; Iruka, Gardner-Neblett, Matthews, & Winn, 2014; Kildare & Middlemiss, 2017; Lê et al., 2013; Libertus & Landa, 2013; Mahalingaiah, Winter, & Aschengrau, 2016; McCormick et al., 2013; Medeiros, Cress, & Lambert, 2016; Morgan et al., 2015; Pasick et al., 2014; Pluymen et al., 2016; Russell, Worsley, & Campbell, 2015; Sansavini et al., 2014; Singh et al., 2014; Sisson et al., 2016; Speirs et al., 2016; Vandecandelaere et al., 2016; Waldie et al., 2014; Walker et al., 2013; H. Wang et al., 2016; M. Wang & Saudino, 2013; Warady et al., 2015; Wimmer, Rothweiler, & Penke, 2017; Wood et al., 2017; Yang & Yang, 2015; Yuan et al., 2016; Zampoli, Pillay, Carrara, Zar, & Morrow, 2016; Zare et al., 2016; X. Zhang & Lin, 2015; Y. Zhang et al., 2017)
Lack of Longitudinal Data	(Buckley et al., 2015; Caemmerer & Keith, 2015; Davies & Oliver, 2016; De Luca et al., 2017; Dussel et al., 2015; Girard et al., 2017; Grabenhenrich et al., 2014; Jeon et al., 2013; Kildare & Middlemiss, 2017; Kohli, Sullivan, Sadeh, & Zopluoglu, 2015; Kremer et al., 2016; Kroksmark, Stridh, & Ekström, 2017; Lê et al., 2013; Lewis et al., 2016; C.-W. Liu et al., 2017; Morgan, Farkas, Hillemeier, & Maczuga, 2016; Morgan et al., 2015; Rachmi, Agho, Li, & Baur, 2017; Royal-Thomas et al., 2015; Rzehak et al., 2013; Sadeghi et al., 2013; Tamayo et al., 2015; Taye et al., 2016; Torres et al., 2015; Treiman, Pollo, Cardoso-Martins, & Kessler, 2013; Xu et al., 2016; Yoon, 2017; Yuan et al., 2016; Zajicek-Farber, Mayer, Daughtery, & Rodkey, 2014)
Incomplete Time Points	(Apouey, 2016; Becker et al., 2014; Darrouzet-Nardi et al., 2016; Flouri, Midouhas, & Joshi, 2014; Kremer et al., 2016; Lewis et al., 2016; McCormick et al., 2016; R. Miller, 2017; Yoon, 2017; Zajicek-Farber et al., 2014)
Inability to Draw Casual Conclusion	(Crampton & Yoon, 2016; Faes, Gillis, & Gillis, 2016; Goelman et al., 2014; Heatly et al., 2015; Mika et al., 2016; Speirs et al., 2016; Uzark et al., 2017; H. Wang et al., 2016; Wood et al., 2017; Wright, Sotres-Alvarez, Mendez, & Adair, 2017; Yang & Yang, 2015; Zampoli et al., 2016)
Noise	(S. T. Baker et al., 2016; Bhattacharya et al., 2014; Brooker & Buss, 2014; Cornwell, McAlister, & Polmear-Swendris, 2014; Gao, Li, Xiong, Shen, & Pan, 2013; Shaw et al., 2014; Shklyar, Pasternak, Kapur, Darras, & Rutkove, 2015; Tharayil et al., 2017; Zare et al., 2016)
Outliers	(Agyei et al., 2016; Becker et al., 2014; Brooker & Buss, 2014; Brown et al., 2017; Derrington et al., 2013; Faes et al., 2016; Greenwood et al., 2013; Ladd-Acosta et al., 2016; Long et al., 2017; Meuwissen & Englund, 2016; M. R. Miller, Müller, Giesbrecht, Carpendale, & Kerns, 2013; R. Miller, 2017; Pérez-Moreno, Blanco-Arana, & Bárcena-Martín, 2016; Schleider et al., 2015; Schott et al., 2013; Tamilia et al., 2016; H. Wang et al., 2016; Zwaigenbaum et al., 2014)

1.4 Research Objectives

This research aimed to develop a missing data imputation framework using machine learning prediction techniques. The main research objectives are, as follows:

- To investigate current academic literature of early childhood, data analysis and data types via systematic review protocol (SLR).
- To explore and investigate NCDRC dataset towards understanding the data behavior and requirement analysis of missing data framework
- To analyze and identify the largest continuous applicable records within NCDRC dataset using Multi-Criteria Analysis

- To explore and design a prediction module for missing data towards reconstructing (NCDRC) dataset using soft computing approach
- To examine and validate the proposed prediction model with respect to data integrity
- To test the developed model on real missing NCDRC dataset scenario

1.5 Research Question

- What is the current state of art in regard with the studies of early childhood, data analysis and data types in the academic literature?
- What is the current utilized approaches for handling missing data in the academic literature within the scope of our systematic review settings?
- Is the current data of NCDRC ready for missing data prediction?
- Are the missing data approaches presented in the literature suitable with NCDRC data?
- How accurate is the prediction module that is based on machine learning measures?
- Can machine learning algorithm be used in real case study?

1.6 Research Scope

This research is aimed to investigate early childhood with respect to available dataset in Malaysia, namely National Child Development Research Center (NCDRC). Therefore, few points need to be taken into account as the following:

- This research is aimed for early childhood studies, though there were some discrepancies to best identify this area considering different authors views and different countries definition for this period in terms of years. Investigations were conducted considering the nature of the data available at NCDRC. Therefore, the selection of this research is based on the age between (Infancy – 5 years) and other rare cases where age was not specially presented in form of years; rather it was presented differently such as kindergarten, pre-school.
- As part of this research scope, the measures considered for data preparation visualization, modulation and analysis for the missing data will not rely on statistical settings. We focus on machine learning, and thus, parameters and preparations are to be considered based on machine learning preferences while addressing the data. Some of the parameters cannot be processed with in their current format for instance, Text and mix-characters. Therefore, data should follow the same data type, due to that, data is converted and grouped to suit machine learning preferences.

1.7 Research Significance

The findings of this thesis would redounds to the benefits of different areas related to early childhood and data analysis studies. As for the area of childhood and medical, it contributes towards identifying what are the issues that encounter these studies, so early actions can be taken to address them. As for the other part of data analysis, it discovers how different area like computer science and machine learning prediction can contribute towards completing children missing data. Therefore, when having larger data, more generalized findings and recommendations can be drawn. For more topics and recommendations that play significance in showing this area of science, please see section Motivation 2.4.2.

1.8 Operational Definitions

Some words and definitions might not be totally clear to some readers, and a clarifications for such elements is good to allow the reader to grasp what this words or phrase is intended for. Therefore, this section aims to display and clarify terms and definitions used in this research, all of them are presented in Table 1.2.

Table 1.2
Operational Definition

CH	Conceptual Variable	Operational Definition
	Bias Findings	Any Systematic Error In An Study That Results In An Incorrect Estimate Of The True Effect Of An Exposure On The Outcome Of Interest
	Casual Conclusions	Expectations Of The Findings
	Data Noise	Amount Of Additional Meaningless Information That Is Not Suitable For Analysis
	Early Childhood	Period From Infancy Until Five Years Of Age
	Longitudinal Data	Data Gathered Over A Long Period Of Time.
	Missing Data	No Data Value Is Stored For The Variable In An Observation
1	Missing Data Imputation	The Process Of Replacing Missing Data With Substituted Values
	Outliers	observations that lies an abnormal distance from other values in a random sample from a populatio
	Sample Size	The Act Of Choosing The Number Of Observations Or Replicates To Include In A Study
	Time Points	Periods of time where records of children were recorded regularly
	Module	Any process taken in this thesis whetherein literature, preprocessing or after pre processing
	Framework	The collection of many modules towards a certain goal.
	(PRISMA) Statement	PRISMA is an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses.
	Search Query	A search query or search term is the actual word or string of words that a search engine user types into the search box
2	Inclusion Criteria	characteristics that the downloaded articles must have if they are to be included in this thesis
	Exclusion Criteria	characteristics that disqualify downloaded articles from inclusion in the thesis
	Taxonomy	the process of naming and classifying articles into groups within a larger mapping, according to their similarities and differences

(*Continue*)

Table 1.2 (*Continued*)

CH	Conceptual Variable	Operational Definition
	Alternative	Available options for the decision to be made
3	Attributes	Also referred to as criteria and used interchangeably in the MCDM contex
	Weight	significances of criteria

1.9 Thesis Layout

This thesis consists of eight chapters; chapter one provided a background about the area of early childhood and data analysis measures in addition to data types. After that, a brief about the current gaps concluded by the state of the problem with regards to missing data, research objective, scope and research questions, the rest of the thesis is organized as the following:

Chapter Two: In Chapter Two, in-depth investigation was conducted for the early childhood studies. This includes defining the terms (Queries) used for investigating the current literature. An (SLR) systematic literature review protocol is adapted to review and analyses the literature towards constructing taxonomy. Articles selected were distributed to map out this area of science and extract important elements like challenges which later on allow us to draw our gaps and research problem.

Chapter Three: In Chapter Three, an overview of (NCDRC) National Child Development Research Center) in Malaysia and brief history about the center. In addition, we highlight this center related reports, data types and funds granted since we are using their data set in this thesis. In addition, last point in this chapter identifies the missing data in the dataset and how it affect it

Chapter Four: in this chapter, the research methodology and the flow of the research are deigned and reported. In addition to that, the main experiments to achieve the research objectives are designed. This includes experiments for data preparation. The main purpose of the data preparation is to transfer the data which is not suited for analysis into a form that unite all the data types in terms of its applicability for analysis. This data is then, grouped by unifying the different ones into one type and make it ready for analysis across groups.

Chapter Five: in this chapter, data modulation and preparation phases starts by navigating data, and describing its parameters and completed statistics, in addition to cleansing all parameters from noise and outliers, and finishing by listing all completed statistics for across different time points in ascending and descending matter to ensure their maximum number.

Chapter Six: in this chapter, all the completed parameters statistics from previous chapter statistics are introduced as alternatives and best one is identified via multi criteria decision making analysis. After that, it is analyzed for correlation, and introduced to next step where missing making scenarios are created with different settings. After the imputation of all scenarios, all the results are compared with original data before missing making to identify their significance differences with the use of three different tests to ensure that best settings and machine learning algorithms for imputations is selected.

Chapter Seven: this chapter includes the selection of actual missing case study data from the NCDRC dataset with settings identified from previous chapter in order to guarantees best results. After selection of best scenario where a good portion of data could be imputed (*n=399*), best performing ML would be utilized to impute it considering it performed best compared with its peers in previous chapter. Last is to measure correlation and significance difference level after the imputation in order to determine if there were significance changes. In addition, a conclusion summary of this entire dissertations including how objectives were achieved across the dissertation, future works, and limitations.