

AN IMPROVED K-NEAREST NEIGHBOR WITH
GRASSHOPPER OPTIMIZATION ALGORITHM FOR
MISSING DATA IMPUTATION

BY

NADZURAH ZAINAL ABIDIN

A thesis submitted in fulfilment of the requirement for the
degree of Master of Computer Science

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia

JANUARY 2020

ABSTRACT

Concurrent with the advanced of data cleaning process, missing data have been influentially known as one of the most common issues encountered for many research area. A real collected dataset such as medical, business, transportation and education are prone to be incomplete or missing especially when the respondents does not respond due to stress, fatigue or inadequacy of knowledge, some of the questions given are sensitive, and lack of option answers presented. One of the mechanisms in solving missing data is through imputation, which is the activity of substituting missing values with plausible records that yield to reasonable accuracy against actual values. A huge number of imputation algorithm has been proposed to estimate the missing values. Unfortunately, most imputation method employed provide less reliable estimations for missing data. Therefore, to accurately deal with missing data, an optimization of one of the state-of-the-art imputation algorithm, K-nearest neighbors (KNN), are proposed to impute those missing values. KNN algorithm has been widely adopted as an imputation algorithm for missing data due to its robustness and simplicity and it is also a promising method to outperform other machine learning methods. However, in many cases, KNN suffers from high computational cost, greater storage requirements, sensitive to noise, high time complexity, and difficult to choose the right centroid position and choice of different function for measuring the distance. Therefore, a conventional way of KNN computes an imputation method still imposes undesirable results. Accordingly, this thesis proposes to develop an optimized KNN imputation method with Grasshopper optimization algorithm (GOA) to present a better imputation result. Grasshopper optimization algorithm is a recent population based metaheuristics which have shown an improved results and efficiencies in tackling issues with missing data. The GOA is incorporated in the algorithm structure, inspired from the natural behavior of grasshopper that maximizes the imputation performance of KNN. The performances of the proposed algorithm will be applied to nine different datasets and compared with other optimization algorithms: Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Dragonfly Optimization (DA), Firefly Algorithm (FFA), Ant Lion Optimization (ALO), and Moth Flame Optimization (MFO), in terms of statistical correlation, error accuracy, and running time. The results show KNNGOA has the most promising performance and outperform among other optimization algorithms with regards to imputation accuracy and fastest time computing for datasets that are large and higher percentage in missing rates (20 percent and above). The analysis of statistical test is also conducted which supports the conclusion of the experiment.

خلاصة البحث

بالتزامن مع التقدم في عملية تنقيب البيانات ، تعتبر البيانات المفقودة واحده من أكثر المشكلات شيوعاً التي واجهتها العديد من مجالات البحث .من المحتمل أن تكون مجموعة البيانات الحقيقية كالبيانات الطبية ، والأعمال التجارية ، والنقل ، والتعليم غير مكتملة أو مفقودة ، وذلك بسبب الإجهاد أو التعب أو نقص في المعرفة ، وايضا تساهم عدم الاجابة على بعض الاسئلة في جعل البيانات ناقصة. يعتبر التضمين إحدى آليات حل البيانات المفقودة. وتتم من خلال التعويض عن القيم المفقودة بقيم معقولة الدقة. و لقد تم اقتراح عدد كبير جدا من خوارزميات التخمين لتقدير القيم المفقودة. لسوء الحظ ، ان معظم طرق التقدير المستخدمة تُقدم تقديرات أقل موثوقية للبيانات المفقودة .لذلك ، للتعامل بدقة مع البيانات المفقودة ، يُقترح تحسين إحدى الخوارزميات وهي خوارزمية الجار الأقرب K Nearest Neighbor algorithm (KNN) لتقدير القيم. تُستخدم خوارزمية KNN على نطاق واسع نظراً لقوتها وبساطتها ، كما أنها تتفوق على أساليب التعلم الآلي الأخرى .ومع ذلك ، في كثير من الحالات ، تعاني KNN من تكلفة حسابية عالية ، متطلبات تخزين أكبر ، حساسية للأخطاء، تحتاج الى الكثير من الوقت وصعوبة اختيار موقع centroid المناسب ، وصعوبة اختيار طريقة لقياس المسافة .لذلك ، فإن الطريقة التقليدية لاستخدام KNN في عملية التضمين لا تزال تؤدي الى نتائج غير مرغوب فيها. وفقاً لذلك ، يقترح هذا البحث تطوير خوارزمية KNN مع خوارزمية Grasshopper Optimization Algorithm (GOA) خلال خوارزمية (KNNGOA) لتقديم نتيجة أفضل لعملية التخمين .تعد خوارزمية (GOA) طريقة تحسين جديدة وقد أظهرت نتائج وفعالية محسنة في معالجة المشكلات المتعلقة بالبيانات المفقودة .وهي مستوحاة من السلوك الطبيعي للجندب . سيتم تطبيق أداء الخوارزمية المقترحة على تسع مجموعات بيانات مختلفة ومقارنتها بخوارزميات التحسين الأخرى مثل : Particle Swarm Optimization (PSO)، Genetic Algorithm (GA)، Dragonfly Optimization (ALO)، Ant Lion Optimization (ALO)، Firefly Algorithm (FFA)، (DA) Moth , (MFO) Flame Optimization (MFO) من حيث الارتباط الإحصائي ، دقة الخطأ ، ووقت التشغيل . وقد أظهرت النتائج أن KNNGOA لديه الأداء المتفوق بين خوارزميات التحسين الأخرى فيما يتعلق بدقة التضمين وأسرع حوسبة الوقت لمجموعات البيانات التي تعد كبيرة وأعلى نسبة مئوية في المعدلات المفقودة (20 بالمائة وما فوق) .يتم إجراء التحليل الإحصائي أيضاً والذي يدعم التجربة.

APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Computer Science

.....
Amelia Ritahani Ismail
Supervisor

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Computer Science

.....
Maziah Abdul Aziz
Internal Examiner

.....
Mohammad Faidzul Nasrudin
External Examiner

This thesis was submitted to the Department of Computer Science and is accepted as a fulfilment of the requirement for the degree of Master of Computer Science.

.....
Raini Binti Hassan
Head, Department of Computer
Science

This thesis was submitted to the Kulliyah of Information and Communication Technology and is accepted as a fulfilment of the requirement for the degree of Master of Computer Science.

.....
Abdul Wahab Abdul Rahman
Dean, Kulliyah of Information
and Communication Technology

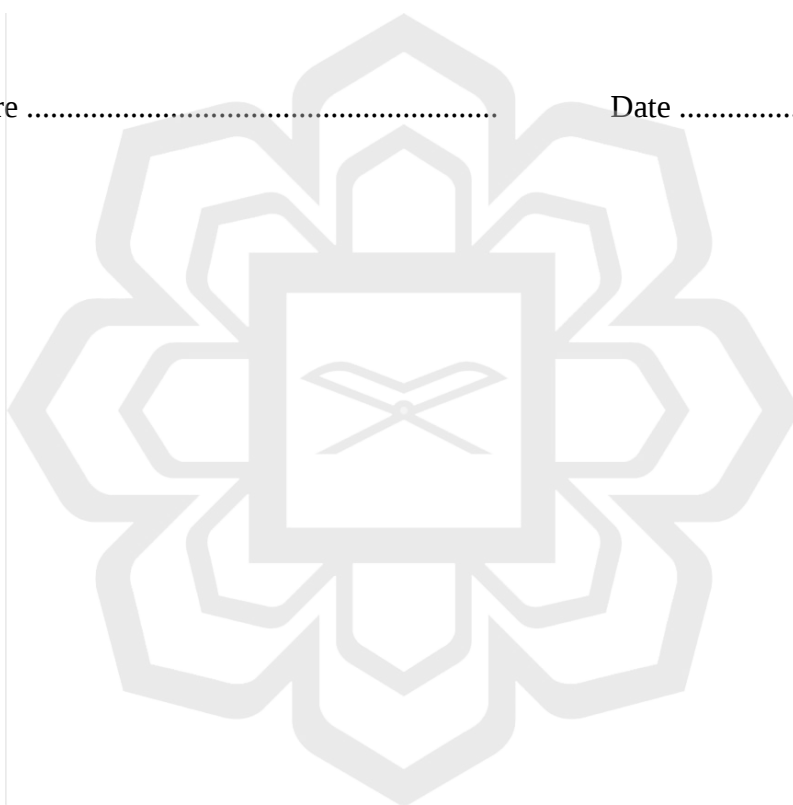
DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Nadzurah Zainal Abidin

Signature

Date



INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

**DECLARATION OF COPYRIGHT AND AFFIRMATION OF
FAIR USE OF UNPUBLISHED RESEARCH**

**AN IMPROVED K-NEAREST NEIGHBOR WITH
GRASSHOPPER OPTIMIZATION ALGORITHM FOR
MISSING DATA IMPUTATION**

I declare that the copyright holders of this thesis are jointly owned by the student and IIUM.

Copyright © 2020 Nadzurah Zainal Abidin and International Islamic University Malaysia. All rights reserved.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.
2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.
3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Nadzurah Zainal Abidin

.....
Signature

.....
Date

ACKNOWLEDGEMENTS

All praises to Allah for the strengths and His blessings for the completion of this research and throughout the duration of my research. I revere the patronage and moral support extended with love, by my parents whose financial support and passionate encouragement made it possible for me to complete this research.

I submit my heartiest gratitude to my respected supervisor, Dr Amelia Ritahani for her constant support, guidance and patience. She taught me how to navigate my journey in the academic world which in turn chart my own personal growth. From the exterior, it might seem I have only learned how to read and write journal papers and eventually this thesis, but I also learned to narrow down my focus for my thesis and discipline myself to finally get the thesis done. Due to her guidance, I've also learned to improve my way of thinking and overcome the challenges in my experiment like a researcher as opposed to merely think like a student. Not only that, I appreciate for her emotional support throughout the process of working on this thesis. This thesis would not have been possible without her timely suggestions despite her busy schedule. May Allah rewards her abundantly.

Above ground, I am deeply indebted to my friends for all the moral support and encouragement throughout the masters' journey. This piece of work I dedicate to my parents, sisters, relatives and friends for their emotional support and for making me believe in myself to keep going and complete this thesis. Thank you for reminding me the importance of working hard and working sincerely.

TABLE OF CONTENTS

Abstract.....	ii
Abstract in Arabic	iii
Approval Page	iv
Declaration	v
Copyright Page	vi
Acknowledgements.....	vii
Table of contents.....	viii
List of Tables	xi
List of Figures	xii
List of Algorithms	xii
List of Abbreviations	xiv
 CHAPTER ONE: INTRODUCTION.....	 1
1.1 Background of study	1
1.2 Problem Background	5
1.3 Research Objectives.....	6
1.4 Research Questions.....	7
1.5 Research Hypothesis.....	7
1.6 Significance and Contribution	8
1.7 Thesis Structure	8
 CHAPTER TWO: LITERATURE REVIEW.....	 10
2.1 Missing Data.....	10
2.2 Machine learning imputation algorithm.....	12
2.2.1 K-nearest neighbors (KNN).....	13
2.2.2 Decision Tree	16
2.2.3 Bayesian Network.....	17
2.2.4 Support Vector Machine (SVM).....	18
2.3 Taxonomy of Missing Data.....	20
2.3.1 Optimizing Imputation Algorithm Approach	21
2.4 Metaheuristics Algorithm	22
2.4.1 Grasshopper Optimization Algorithm (GOA)	23
2.5 Literature Summary	29
 CHAPTER THREE: RESEARCH METHODOLOGY	 32
3.1 Research Methodology	32
3.1.1 Phase One – Literature Review.....	33
3.1.2 Phase Two – Data Acquisition.....	34
3.1.3 Phase Three – Experimental Setup	35
3.1.4 Phase Four – Performance Analysis	36

3.1.5 Phase Five – Documentation	37
3.2 Mapping research methodology, research objectives, and research questions	37
3.3 Chapter Summary	38
CHAPTER FOUR: K-NEAREST NEIGHBORS BASED ON GRASSHOPPER OPTIMIZATION ALGORITHM (GOA)	39
4.1 Optimized Imputation Algorithm	39
4.2 Optimized K-nearest neighbors with Grasshopper Optimization Algorithm.....	41
4.3 Experimental Setup	45
4.4 Experimental Data	47
4.4.1 Chronic Kidney Disease (CKD) – KD1 & KD2.....	49
4.4.2 Hepatocellular Carcinoma (HCC) Survival.....	50
4.4.3 Acute Kidney Injury (AKI).....	50
4.4.4 Environmental Health Perspectives (EHP) Phthalates	51
4.4.5 Electrocardiogram (ECG) Heartbeat.....	51
4.4.6 Blood test analysis	52
4.4.7 Automobile	52
4.4.8 Air Quality	52
4.5 Evaluation Metrics	53
4.6 Performance Metrics	54
4.6.1 Statistical Correlation.....	54
4.6.2 Error Accuracy.....	55
4.6.3 Running Time	56
4.7 Statistical Significance Test.....	56
4.8 Chapter summary	58
CHAPTER FIVE: RESULTS AND DISCUSSION.....	59
5.1 KNN and GOA Experimental Results	59
5.2 Performance Metrics Results	66
5.2.1 Statistical Correlation.....	66
5.2.2 Error Accuracy.....	83
5.2.3 Running time.....	86
5.3 Statistical Significance Test.....	88
5.4 Chapter summary	93
CHAPTER SIX: CONCLUSION AND FUTURE WORKS	95
6.1 Conclusion	95
6.2 Research Objective Achievements	98
6.3 Future Work.....	99
REFERENCES.....	101

APPENDIX A: ALGORITHMS' PARAMETERS SETTINGS.....109



LIST OF TABLES

<u>Table No.</u>		<u>Page No.</u>
2.1	Literature Summary	29
4.1	Experimental Data	48
4.2	Significance Table for Vargha Delaney A Test	57
5.1	Results of Error Accuracy for Machine Learning Classifiers	61
5.2	Results from KNN imputation for 10%, 30%, and 50% of missing rates	63
5.3	Error accuracy for GOA based on other machine learning algorithms	65
5.4	Summary Analysis for Statistical Correlation	82
5.5	Error Accuracy for all datasets	82
5.6	Running time for all datasets	87
5.7	The magnitude of difference indicated by Vargha Delaney A Test	89
6.1	Research objective achievements	98
Appendix A	Algorithms Parameter Settings	109

LIST OF FIGURES

<u>Figure No.</u>		<u>Page No.</u>
2.1	Frequency of algorithms from literature studies	13
2.2	Optimal Hyperplane (Village Idiot, 2012)	19
2.3	Taxonomy from literature studies	20
2.4	Life cycle of a grasshopper	24
2.5	Primitive corrective pattern between individual in a swarm of grasshoppers	24
3.1	Research methodology phases	33
3.2	Mapping methodology phases with research objectives and questions	38
4.1	The flowchart for KNNGOA algorithm	44
4.2	Experimental Setup	45
5.1	Scatterplot for KD1 dataset	67
5.2	Scatterplot for KD2 dataset	69
5.3	Scatterplot for HCC Survival dataset	70
5.4	Scatterplot for AKI dataset	72
5.5	Scatterplot for EHP dataset	73
5.6	Scatterplot for ECG Heartbeat dataset	75
5.7	Scatterplot for Blood Test dataset	76
5.8	Scatterplot for Automobile dataset	78
5.9	Scatterplot for Air Quality dataset	79

LIST OF ALGORITHMS

<u>No.</u>		<u>Page No.</u>
1	Grasshopper Optimization Algorithm (GOA)	27
2	K-nearest neighbors Grasshopper Optimization Algorithm (KNNGOA)	42



LIST OF ABBREVIATIONS

KNN	K-nearest neighbors
GOA	Grasshopper Optimization Algorithm
GA	Genetic Algorithm
PSO	Particle Swarm Optimization
DA	Dragonfly Algorithm
FFA	Firefly Algorithm
ALO	Ant Lion optimization
MFO	Moth Flame Optimization
SVM	Support Vector Machine
LDA	Linear Discriminant Analysis
LLS	Local Least Square
PCA	Principal Component Analysis
MV	Missing Value
MAE	Mean Absolute Error
MSE	Mean Square Error
RMSE	Root Mean Square Error
MAPE	Mean Absolute Percentage Error
MCAR	Missing Completely at Random
MAR	Missing at Random
MNAR	Missing not at Random
SVD	Single Value Decomposition
KNNGA	K-nearest neighbors with Genetic Algorithm
KNNPSO	K-nearest neighbors with Particle Swarm Optimization
KNNGOA	K-nearest neighbors with Grasshopper Optimization Algorithm

KNNDA K-nearest neighbors with Dragonfly Algorithm
KNNFFA K-nearest neighbors with Firefly Algorithm
KNNALO K-nearest neighbors with Ant Lion Optimization
KNNMFO K-nearest neighbors with Moth Flame Optimization



CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF STUDY

Data is an essential asset for any discipline of work to efficiently analyze a better decision making and it is accessible in every edge of life which provide different insights. The first step in data mining, with respect in collecting data, a researcher must confront with common problems that any data are prone to. Practically, data collected that inclined to missing is a major source of poor data quality. Poor data quality is inevitable which could lead your analysis utterly wrong and adversely affect the mining results regardless data mining algorithm (Mouroutis, 2015; G. Rahman & Islam, 2013; Shahzad & Ahmed, 2017).

Approximately, 67.6 percent of real-world datasets has fully acknowledged missing data problem (Aliya Aleryani, Wenjia Wang, & Beatriz De La Iglesia, 2018; Dong & Peng, 2013; Kang, 2013; Kenward, 2013), although there are increasing evidence that give attention to detail on the solution. Besides, more than 40 percent of datasets embedded in UCI Machine Learning Repository which is a platform that extensively being refferred to make an empirical analysis are missing (Pan, Yang, Cao, Lu, & Zhang, 2015). In general, there are many reasonable justifications that contribute to the huge percentage of missing data which includes human error due to stress or fatigue or inadequacy of knowledge, refusal to answer questions especially when it is related to privacy issues, do not understand questions given, respondents migration, early successes of a treatment, and treatment or instrumental failures, instrument error, adverse events, and death (De Leeuw, Hox, & Huisman, 2003; Hand, 2018; Hughes,

Heron, Sterne, & Tilling, 2019; Kenward, 2013; Ramli, Yahaya, Ramli, Yusof, & Abdullah, 2013; SPSS White Paper, 2009).

Data cleaning is an essential process to improve efficiency of analyzing data and to ensure the quality. Data cleaning is a process of detecting and removing errors and incompleteness from data in order to improve the quality of data (Rahm & Hong Hai Do, 2000). Most healthcare, business, education, society and transportation datasets were found to be incomplete, which double suffers to perform task of data mining. This is due to the fact that incorrect prediction measures may leads to improper medical treatment, incorrect inference for business decision, wrongly grade predictions by teachers, stranded community, and under-provide accessibility for transportation (Khaldy & Kambhampati, n.d.; Lin, 2005; Tofallis, 2015). One of the major task in data cleaning phase is to fill missing values.

Many investigation towards techniques has been conducted as the solution of missing datasets, in particular case deletion, mean substitution and imputation. Case deletion plainly discard the incomplete cases from the original datasets which yield to many serious problem than solutions because the original dataset most likely differ from the group with missing data (M. Rahman & Davis, 2013). Mean substitution approach used mean value of a variable in place of the missing value for the same variable which can cause bias and lead to invalid conclusions (Kang, 2013). While imputation refers to a routine of substituting missing values with plausible values that approximately resemble the actual values. Among all three mentioned treatment towards incomplete dataset, the most adequate solution is an imputation. (Usharani & Sammulal, 2016) also reported that there are seven research issues when handling with datasets, which includes the imputation of missing values (A. Wang, Lim, Cheng, & Xie, 2018). Research on imputation methods can be discovered through statistical and machine

learning algorithm. From statistical standpoint, missing value can increase the chance to lead a massive loss information, biased parameter estimates, and reduce statistical power (Dong & Peng, 2013; Nakagawa, 2015; Rosenthal, 2017). Therefore, the advantages and limitations of statistical algorithm has been adopted into machine learning algorithm to perform better approaches in imputing missing values due to have greater performance in imputation compared to statistical methods (Jerez et al., 2010; M. Rahman & Davis, 2013).

A series of studies have been proposed machine learning as an imputation algorithm, and yet, there is no imputation algorithm that consistently outperforms others in every situation (Liu & Gopalakrishnan, 2017). However, selecting the most appropriate algorithm may significantly improve the accuracy of imputation results (Liew, Law, & Yan, 2011). Among all machine learning imputation algorithms, KNN algorithm has been widely adopted as an imputation for missing data and it is also a promising method to outperform other machine learning methods. KNN is a straightforward, yet powerful classification algorithm that computes a values estimates from the closest neighbors which has relatively high accuracy (Alizadeh, Minaei-bidgoli, & Amirgholipour, 2009; Batista & Monard, 2002; Pan et al., 2015). The favorable points on KNN are simplicity, comprehensibility and scalability (Iren & Tokle, 2017; Lamba & Kumar, 2016). Despite the simplicity associated with KNN algorithm, several studies have well acknowledged that KNN suffers from high computational cost, greater storage requirements, and sensitivity to noise. Therefore, a conventional way of KNN computes an imputation method still imposes undesirable results.

Therefore, this research propose to optimize KNN imputation method with Grasshopper optimization algorithm (GOA) to improve its performance. This

optimization algorithm mimics the behavior of grasshopper swarm for solving the optimization problems (Saremi, Mirjalili, & Lewis, 2017). GOA is a recent swarm intelligence algorithm that was developed by Mirjalili et al. in 2017 and used population based method. A life cycle of grasshopper passes through three main stages; egg, nymph and adult. The nymph grasshopper does not have wings, thus they move slowly to eat all vegetation on their path (Mafarja et al., 2017). However, after a period of time, grasshopper will become an adult with wings to form a swarm in the air and move fast to large scale region (Neve, Kakandikar, & Kulkarni, 2017). For the purpose of this study, GOA favors KNN imputation methods by surviving to avoid local optima and able to find the global space in the given space. Nevertheless, GOA beneficially balance an exploration and exploitation to drive grasshopper towards global optimum. GOA algorithm presents an accurate solution and results for constrained and unconstrained optimization problem respectively (Luo, Chen, Xu, Huang, & Zhao, 2018).

A fundamental assumptions of GOA that may improve the processes of KNN imputation can be found in the way GOA find its optimum solution. KNN estimates a values from its nearest neighbors while GOA have a high avoidance to find a solution between a set of neighborhood and provides solution among all possible solution. Besides, one of the limitations of KNN imputation is the algorithm searches through all datasets for estimating most similar instances, which takes a great deal of time. GOA favors KNN imputation in the sense of running time where one of the main characteristics of grasshopper in adulthood phase is long range and abrupt movement.

1.2 PROBLEM BACKGROUND

K-nearest neighbors (KNN) algorithm is one of the promising methods that may assist a decision maker to help making a better prediction especially in dealing with missing value (Jabbar, Deekshatulu, & Chandra, 2013). KNN is a straightforward, yet powerful classification algorithm that computes a values estimates from the closest neighbors. The favorable points on KNN are simplicity, comprehensibility and scalability (Lamba & Kumar, 2016). Despite the simplicity associated with KNN algorithm, there are numerous uncertainty to the prediction measures (Luis & Medina, 2013), which degrading the accuracy by the presence of missing value (Wu et al., 2008). Several studies have well acknowledged that KNN suffers from high computational cost, greater storage requirements, and sensitivity to noise (Batista & Monard, 2002; Classifier, Prasath, Arafat, Alfeilat, & Lasassmeh, 2017; X.-X. Wang & Ma, 2014). For the first two problems, previous studies had proposed different solutions depending on reducing the size of the training dataset (Alizadeh et al., 2009; Manning, Raghavan, & Schutze, 2008; Zheng, Guo, Tung, & Wu, 2016). The third problem was solved by considering to selectively choose k -value and distance metrics, which primarily determined the performance of KNN. Although for all the problems aforementioned above has been investigated and solved, no previous studies that exclusively focused to optimize the performance of KNN imputation algorithm using optimization algorithm except (De Silva & Perera, 2016).

The plausible solution on optimizing KNN imputation approach proposed by (De Silva & Perera, 2016) presents an integration between KNN and Genetic Algorithm (GA) as an imputation algorithm. The authors attempt to provide a comparison of proposed solution with constrainedly among conventional KNN imputation algorithm

and other basic imputation methods such as mean imputation and median imputation. The findings of the study assessed the performance of proposed solution is inconsistent. The proposed solution merely performs better at certain high level of missing rate for large dataset than a small missing rate in a dataset.

One way of recovering from this problem is to address the KNN imputation methods with more sophisticated algorithm. This sophisticated algorithm is proposed by integrating KNN with nature-inspired optimization algorithm, Grasshopper Optimization Algorithm (GOA). Many aspects inspired from Grasshopper Optimization Algorithm as regards to their swarming behavior, biological behaviors, and surviving strategy (Luo et al., 2018; Neve et al., 2017; Sulaiman et al., 2019). Grasshoppers are powerful insects with long and large hind jumping legs as a booster to propel in the air, where they spread the wings and take off. These creatures also have a specific social network which equip them with a particular predatory strategy. All these special characteristics of grasshopper are believe to overpower the KNN imputation methods to estimates the missing values.

1.3 RESEARCH OBJECTIVES

This research work sought to achieve the following objectives:

1. To review the existing imputation methods using machine learning algorithms.
2. To develop an optimized KNN imputation method with Grasshopper Optimization Algorithm.
3. To evaluate the performance of an optimized KNN imputation method with Grasshopper Optimization Algorithm.

4. To validate the performance of optimized KNN imputation method with Grasshopper Optimization Algorithm with other optimization algorithm with KNN imputation algorithm.

1.4 RESEARCH QUESTIONS

This study attempts to answer the following research questions:

1. What are the machine learning imputation algorithms that have been developed?
2. How does the Grasshopper Optimization Algorithm improve KNN imputation?
3. How effective is the proposed solution compared to other optimization imputation methods?

1.5 RESEARCH HYPOTHESIS

A general hypothesis for this research is imputation of missing values can be formulated and established using KNN with Grasshopper Optimization Algorithm (GOA), which overcomes the issues of the conventional techniques of KNN to impute missing values.

Below is the hypothesis of the study:

H₁: The optimized KNN with Grasshopper Optimization Algorithm (GOA) has shown good performance in terms of the performance metrics and statistical analysis as compared to other optimized KNN based on optimization algorithm when imputing missing values.

A comprehensive hypotheses on statistical analysis is conversed in Section 5.3.

1.6 SIGNIFICANCE AND CONTRIBUTION

This thesis contributes to knowledge by firstly, conducting a review and provide a comparative study of imputation algorithms. The significance of the difference in performance of the methods is presented. Secondly, the significant contribution of this thesis is to propose an optimized KNN imputation algorithm with grasshopper optimization algorithm (GOA) for missing value problems.

1.7 THESIS STRUCTURE

The thesis is broken down into 6 chapters where we describe the enhancement process of KNN imputation method to achieve the highest accuracy results for missing data problem.

In Chapter 2, we provide a review of the related work and literature on missing data and imputation method. This chapter attempt to review and clarify the four well-established missing data treatment (k-nearest neighbors, decision tree, Bayesian network, and support vector machine) discovered by machine learning. Secondly, the chapter introduce a literature analysis that relevant to fill the gap and further investigate the imputation techniques.

In Chapter 3, we provide the research methodology that describes the work to introduce KNN based on seven optimization algorithm such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Grasshopper Optimization Algorithm (GOA), Dragonfly Algorithm (DA), Firefly Algorithm (FFA), Ant lion optimization (ALO), and Moth Flame Optimization (MFO).

In chapter 4, we detail out our attempt to develop an optimized KNN imputation algorithm, KNNGOA, its pseudo code, and steps. Chapter 5 explains the experimental setup of the research, experimental data, metrics evaluation and performance metrics that evaluate the model KNNGOA algorithm with other seven optimization algorithm.

Chapter 6 focus on explaining the simulation results that we obtained from the KNN imputation methods based on seven optimization algorithm. Finally, Chapter 7 is eluciated the work of the thesis that reflect on the conclusion and future works based on the KNN imputation method based on GOA.

