

University of Nottingham
UK | CHINA | MALAYSIA

A HYBRID GENETIC ALGORITHM BASED APPROXIMATE CASH CROP MODEL WITH SUPPORT VECTOR MACHINE CLASSIFIER FRAMEWORK FOR PREDICTING ECONOMIC VIABILITY OF UNDERUTILISED CROP IN RURAL AREA

Oh Mei Shin

10/23/2017

Thesis submitted to The University of Nottingham for the Degree of Doctor of Philosophy

Abstract

The research deals with developing a hybrid Genetic Algorithm based on approximate cash crop model framework to predict economic viability of underutilised crop in rural area using Support Vector Machine. Machine learning techniques are dependent on large amount of data in order to predict outcomes. However, underutilised crops are, by definition, not cultivated on a large commercial scale causes the scarcity of training data for this application. In this proposed framework, SVM is implemented in conjunction with a Genetic Algorithm (GA) and associated fitness functions to generate training data for the SVM from approximate models developed for normal cash crops. Approximate models are used in Genetic Algorithm as fitness function to generate synthetic data. Synthetic data generated is used to train Support Vector Machine. Experiments are designed to compare synthetic data from actual models and data generated from approximate models. Model and real data from World Bank is also used to validate the proposed framework. Finally, synthetic data generated from approximate farm income models for crop are tested against the real village data obtained from Crops For the Future Research Centre. The result shows that good classification is attainable in spite of the inaccuracy that existed in the training data from an approximate model and artificially generated through Genetic Algorithm which includes constraints that reflect physical conditions found in rural villages. This framework is able to identify village which is able to achieve potential success economically planting underutilised crops before cultivating the crop itself based on the approximate model from village who has successfully commercialised normal crops.

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my main supervisor, Dr. Chen ZhiYuan for her continuous support throughout my research with patience and invaluable guidance. She has provided me with opportunity to overcome my limits.

I am extremely thankful to my second supervisor, Prof. Dino Isa who is tremendously passionate in this research area and provided me with invaluable insights on this research work. I am grateful for his continuous motivation and encouragement throughout my PhD journey.

A special gratitude to Crops For the Future Research Centre (CFF) for the financial support to complete my research and provided an amazing experience to work with researchers from transdisciplinary area.

I am indebted to Dr. Timothy Brailsford and his wife, Dr. Susan Scarborough for their generosity hosting me when I was in Nottingham, UK as part of the PhD journey. Dr.Tim has also provided me with constructive comments which helped me to improve my thesis.

I am grateful to have my beloved family and friends for their support spiritually and their faith in me along the journey. I thank my fellow lab mates in BB79e for stimulating discussion, also organizing occasional gatherings which opened my eyes and mind to different cultures.

Finally, I would like to express my deepest appreciation to my mother who sacrificed so much to give me the best and provided me with mental and emotional support unconditionally which keeps me motivated.

Contents

Abstract i

Acknowledgement ii

List of Figures vii

List of Tables ix

Chapter 1 Introduction 1

1.1 Research Overview 1

1.2 Problem Statement 4

1.3 Research Issues 4

1.4 Aims and Objectives 6

1.5 Scope of Research 7

1.6 Contributions 8

1.7 Organization of Thesis 8

Chapter 2 Literature Review 10

2.1 Introduction 10

2.2 Importance of Underutilised Crops 11

2.3 Commercialisation of Crops 12

2.3.1 Promoting Commercialisation of Underutilised Crops 15

2.4 Review of Methods used to generate Artificial Data 22

2.4.1 Data Generation from distribution function of original samples 23

2.4.2 Data Generation from original data perturbation 27

2.4.3 Data Generation from other methods 31

2.4.4 Analysis on Methods of Artificial Data Generation 34

2.5 Data Labelling 37

2.6 Economic Models 37

2.6.1 Economic Models based on Mathematical Approach 38

2.6.2 Economic Models based on Surveys (Empirical Approach) 40

2.6.3 Economic Models based on Simulation 42

2.6.4 Other Economic Models 44
2.6.5 Limitations and Inaccuracies of Economic Models 45
2.6.6 Machine Learning vs Economic Modelling 47
2.6.7 Review on Methods of Improving Traditional Modelling Methods 50
2.6.8 Substitution of Machine Learning Models to Traditional Models 50
2.6.9 Integration of Machine Learning Models to Traditional Models 51
2.6.10 Analysis on Methods of Improving on Traditional Modelling Methods 51
Chapter 3 Relevance of Support Vector Machine and Genetic Algorithm 53
3.1 Introduction 53
3.2 Genetic Algorithm 53
3.2.1 Schema Theorem 54
3.2.2 Modification from Schema Theorem 59
3.2.3 Section Summary 63
3.3 Statistical Learning Theory 64
3.3.1 Empirical Risk Minimization 65
3.3.2 VC Dimension 66
3.3.3 Structural Risk Minimization 68
3.3.4 Support Vector Machines 68
3.3.5 Quadratic Programming 73
3.3.6 Section Summary 75
Chapter 4 Methodology 77
4.1 Introduction 77
4.2 Economic models 80
4.2.1 Farm Income Models 81
4.2.2 Analysis on Economic Models 87
4.2.3 Analysis on Farm Income Models 89
4.2.4 Inputs variables to the models 92

4.2.5 Constraints 93
4.2.6 Summary of Variables 100
4.3 Artificial Training Data Generated by Genetic Algorithm 101
4.4 Data Labelling 106
4.5 Kernel Trick 108
4.5.1 Mercer's Theorem 109
4.5.2 Wavelet Kernel 110
4.5.3 Making New Kernels from Kernels 118
4.5.4 Hybrid Kernel for SVM 118
4.5.5 A New Kernel – Champernowne Kernel 119
4.6 Support Vector Machine Implementation 121
4.7 System Implementation 122
4.7.1 Real Application 124
Chapter 5 Results and Discussions 125
5.1 Introduction 125
5.2 Performance Measure 125
5.3 Robustness of SVM on Approximate Models 126
5.3.1 Approximate Model 1 126
5.3.2 Approximate Model 2 129
5.3.3 Section Analysis 132
5.4 Comparison of new kernels on benchmark data 133
5.4.1 Section Analysis 134
5.5 Feasibility on other applications and proximity of synthetic data generated to the real data 134
5.5.1 World Bank GNI Data with Synthetic Data Training and Test 137
5.5.2 Comparison of Proposed Methodology with Other Methods 138
5.5.3 Section Analysis 142
5.6 Parameters Adjustment in Genetic Algorithm 143

5.6.1 Parameter 1: Population Size 144
5.6.2 Parameter 2: Initial Population Range 149
5.7 Real Village Data Validation 150
5.8 Summary 158
Chapter 6 Conclusion and Future Works 159
6.1 Conclusion 159
6.2 Future Works 161
References 163
Appendix A: Report on Using Synthetic Data Generated from GA for Support Vector Regression Training to Predict Cost of Production of Underutilised Crop 197

List of Figures

Figure 1.1 Examples of Underutilised Crops: (a) Bambara Groundnuts (b) Fonio [4][5] 2
Figure 2.1 Pie Chart of Food Energy Supply for Major Staple Crops 11
Figure 2.2 Commercialisation Process [20] 13
Figure 2.3 Agricultural Commercialisation Process [21] 14
Figure 2.4 Process of Commercialisation Underutilised Crops [25] 21
Figure 2.5 (a) Original Data (Top), (b) Transformed Data (Bottom Left), (c) Acyclic model that indicates the path to travel (Bottom Right) [35] 26
Figure 2.6 Interpolation of neighbour lines from data point $\boldsymbol{xi}$ [42] 29
Figure 2.7 Case Study on evaluating the risk level of customer leaving the service of a telecom company [107] 49
Figure 3.1 Punctuated Equilibrium of fossils [123] 63
Figure 3.2 Concept of Structural Risk Minimization [126] 67
Figure 3.3 Which hyperplane is the optimal?[129] 70
Figure 3.4 Optimal Hyperplane [129] 70
Figure 3.5 Function of Hyperplane[130] 71
Figure 4.1 Genetic Algorithm Artificial Data Generation Process 101
Figure 4.2 Example of Data Generated from Genetic Algorithm 104
Figure 4.3 Breakpoint Label Example 1 107
Figure 4.4 Breakpoint Label Example 2 107
Figure 4.5 Mapping nonlinearly separable data to higher space [185] 108
Figure 4.6 Path of contour integration [189] 115
Figure 4.7 Champernowne distribution function 120
Figure 4.8 SVM Implementation Phases 121
Figure 4.9 Overall System Architecture 123
Figure 5.1 Data Summary Plots of Variables for GNI Model 136
Figure 5.2 Generation versus Fitness Value Plots of Model 1 for Population Size 10 (Top Left), 100 (Top Right), 1000 (Bottom Left) and 10000 (Bottom Right) 145
Figure 5.3 Generation versus Fitness Value Plots of Model 2 for Population Size 10 (Top Left), 100 (Top Right), 1000 (Bottom Left) and 10000 (Bottom Right) 146
Figure 5.4 Generation versus Fitness Value Plots of Model 3 for Population Size 10 (Top Left), 100 (Top Right), 1000 (Bottom Left) and 10000 (Bottom Right) 147
Figure 5.5 Population Size vs Elapsed Time Plots for Model 1(Top Left), 2(Top Right) and 3 (Bottom) 148
Figure 5.6 Balanced Example from Generation versus Fitness Value Plot 149
Figure 5.7 Imbalanced Example from Generation versus Fitness Value Plot 150

Figure 5.8 Predicted Labels of SVM Trained by Data Generated from Model 2 155
Figure 5.9 Predicted Labels of SVM Trained by Data Generated from Model 3 157

List of Tables

Table 3.1 Examples on metastable state of different fitness function..... 61
Table 4.1 Methodology steps to solve Problem Statements..... 78
Table 4.2 Details of Variables Involved in Model 1 83
Table 4.3 Details of Variables Involved in Model 2 84
Table 4.4 Details of Variables Involved in Model 3 85
Table 4.5 Overview of Model 1,2 and 3 87
Table 4.6 Central Tendency Measures 92
Table 4.7 Minimum and Maximum Farm Size in Kenya, Ethiopia, Malawi, Niger, Nigeria, Tanzania, Uganda, Bangladesh, Nepal, Vietnam, Bolivia, Guatemala, Nicaragua and Albania 95
Table 4.8 Market Distance of Villages in Philippines 97
Table 4.9 Market Distance of Villages in India..... 97
Table 4.10 Material Input Costs Per Acre of Rice Production in US 98
Table 4.11 Material Input Costs Per Hectare in Indonesia..... 99
Table 4.12 Summary of Lower and Upper Bounds for Each of the Variables 100
Table 4.13 Kernels..... 110
Table 4.14 Wavelet functions and corresponding mother wavelet and kernel functions 112
Table 5.1 Actual Model vs Approximate Model for Ohm's Law..... 126
Table 5.2 Average Accuracy of Experimental Model 1 (Population Size of 20) 127
Table 5.3 Average Accuracy of Experimental Model 1 (Population Size of 150) 128
Table 5.4 Average Accuracy of Experimental Model 1 (Population Size of 800) 129
Table 5.5 Actual Model vs Approximate Model for Motion..... 130
Table 5.6 Average Accuracy of Experimental Model 2 (Population Size of 20) 130
Table 5.7 Average Accuracy of Experimental Model 2 (Population Size of 150) 131
Table 5.8 Average Accuracy of Experimental Model 2 (Population Size of 800) 131
Table 5.9 Relationship between Models 132
Table 5.10 Benchmark Data Characteristics 133
Table 5.11 Performance Comparison of between Commonly Used Kernels and New Kernels..... 134
Table 5.12 Average Accuracy of GNI Model..... 137
Table 5.13 Gaussian Distribution Synthetic Training Data Generation Performance. 139
Table 5.14 Uniform Distribution Synthetic Training Data Generation Performance .. 140
Table 5.15 Synthetic Training Data Generation Proposed by Yang et al. [28] *Performance*..... 141

Table 5.16 Compare between proposed methodology and other methods 142
Table 5.17 Population Size vs Elapsed Time of Model 1 144
Table 5.18 Population Size vs Elapsed Time of Model 2 146
Table 5.19 Population Size vs Elapsed Time of Model 3 147
Table 5.20 Average Accuracy of Model 2 on Real Village Data 151
Table 5.21 Comparison of Variables Between Real Village Data and Model 1,2 and 3 153
Table 5.22 Average Accuracy of Model 3 on Real Village Data 156

Chapter 1 Introduction

1.1 Research Overview

Machine Learning requires training data in order to predict future outcomes. In this project, the prediction of whether a village can succeed in growing and commercializing underutilised crops is the main aim. However, this effort is hampered by the lack of data in this domain due to the fact that underutilised crops are, by definition, not cultivated on a large commercial scale. It is a contention proposed in this research that this can be overcome by using models available in literature developed for ordinary commercial crops to generate training data with which to train an inference engine. This data, artificially generated through the use of a model acting as the objective function in a Genetic Algorithm, is additionally "inappropriate" because it is not for underutilised crops. However, it is the aim of this research to show that the SVM will still be able to predict the outcome of introducing an underutilised crop to a rural area on the basis that the factors leading to commercialisation do not differ between underutilised and ordinary crops.

As such, the factors that are considered is independent of the plant biology; instead the data does characterize the area and population that had successfully commercialised a certain crop. Underutilised crops are crops that were once used

for food or medicine but gradually became less important due to various factors and constraints. These crops are mostly grown and used by the locals. One of the factors the crops are underutilised is due to the lack of supply by farmers. This situation occurs owing to lack of competitive edge of underutilised crops as compared to other crops like rice and wheat. However, underutilised crops play an important role in helping the poor in creating revenue [1][2][3]. Figure 1.1 shows two examples of underutilised crop, Bambara groundnuts and fonio that can be found in rural areas. Multiple uses offer greater opportunities to raise income of local people by diversifying plant products. This will contribute to improving the well-being of these people in rural areas.



Figure 1.1 Examples of Underutilised Crops: (a) Bambara Groundnuts (b) Fonio [4][5]

The problem needs to be formulated in terms which is suitable for use with machine learning techniques, namely the Support Vector Machine (SVM). As such Support Vector Machine is trained with data for rural villages which is generated from Genetic Algorithm, where the objective function of Genetic Algorithm to generate the data is established from models which reflects the relationship between characteristics of rural villages and their income generated. This is being done because of the lack of data for the different types of rural villages which may be able to cultivate and commercialise underutilised crops. The reason that a machine learning technique is used is because it is able to objectively classify villages based on the potential of success in commercialising underutilised crops. The models that are used as objective function is approximate models of underutilised crops. Generally, an approximate model is used to approximate a complicated as well as computational cost intensive process into a simplified model [6][7]. An approximate model is also used when the information that is provided is incomplete to build an exact model [8]. In the context of this research, an approximate model is referring to a normal crop model that includes characteristics or variables that are approximately similar to those of an underutilised crop model. The approximate models can be used in place of exact underutilised crop model because variables are independent of the crops. Alternatively, the variables of these approximate models are dependent on

the villages. Among some of the significant variables are experience of farming, distance between the village and the nearest market, land size, weather condition and so on. In order to properly formulate the objective function, a literature search was done to find the significant variables and its relationship with one another to identify farm income [9]–[11].

1.2 Problem Statement

95% of food for the world is provided by only 30 species of plants out of 7000 species. Maize, wheat and rice are three major crops that made up to more than 50% of global consumption. In order to ensure biodiversity and future food security, underutilised crops are gaining research attention. However, to transform underutilised crops into economically successful crops, farmers need assurance of successfulness in selling the crops before planting the new crops. Hence, a framework to predict the economic viability of underutilised crops in a village is proposed in this research.

1.3 Research Issues

The research issues are refined as follows:

1. **Decision support systems** for farmers in determining whether or not to cultivate new underutilised crops as opposed to traditional ones are not well developed [12]–[14].

The proposed system predicts the potential as to whether underutilised crops will be successful in bringing economic growth to farmers by using a combined machine learning and modelling method. This is done before having the crops planted physically and will save a tremendous amount of cost and time.

2. **Economic models for underutilised crops are not available**. This is due to the fact that there are not many instances of underutilised crops that have been commercially successful. In this research new models for these crops are presented which are derived from models used for normal commercial crops. The new models are also formed using curve fitting software when small amounts of data are available. However, these models are not accurate.
3. **Inaccuracy of the approximate models** is mitigated by combining the models with a machine learning technique, namely the Support Vector Machine. Inaccuracy is induced in the approximate models as the models are not formulated based on underutilised crops data. Assumptions that was made when formulating the models have also contributed to the inaccuracy. This will be discussed in details in Section 2.6.5. Machine learning technique is able to overcome these by recognising the pattern lies within the data without overfitting the data. In particularly, SVM

avoids overfitting the inaccuracies of data by adjusting the misclassification penalty parameter, C.

4. **The unavailability of a large number of training data** to train Support Vector Machine is an issue which needs to be addressed if machine learning is to be used for accurate classification. The Genetic Algorithm is proposed as a means to generate sufficient artificial training data to train SVM, as constraints can be used to reflect real life conditions and the model can be used as the objective functions.
5. **Proper labelling of training data is also an issue** if classification accuracy is to be maintained above 90% for all possible cases. As such a "natural way" of labelling the data before training the SVM will need to be implemented.

1.4 Aims and Objectives

The aim of this research is to develop a hybrid Genetic Algorithm based approximate cash crop model with Support Vector Machine classifier framework that can be used to predict economic viability of underutilised crop in rural area. Classification is carried out to find out whether underutilised crops will be successful or not in a particular village. The following objectives have to be fulfilled to achieve main aim.

- To use appropriate models that approximate underutilised crops in terms of village characteristics. Farm income models that relate to farm income with village characteristics are suitable to be used in this research as will be illustrated in Chapter 4.
- To apply Support Vector Machine for classification based on data generated from approximate model.
- To implement Genetic Algorithm in order to generate sufficient artificial training data to train SVM.

1.5 Scope of Research

This research does not cover factors related to plant biology, soil requirements, and factors related to plant science per se. This is because it is hypothesized that factors related to the characteristics of the place and the farmers themselves can give a good indication of success without considering the plant itself. After this method is applied, then the plant scientist should test the soil and other factors related directly to the plant to identify the exact crop which can be cultivated. As such with the separation of economic system and plant biology, chances of success is anticipated to increase due to the fact that the systems are independent of each other. This allows the reduction of work for the plant biologists as the

places which can be commercially successful are now reduced to the ones where the place characteristics are most likely to bring commercial success.

1.6 Contributions

This section highlights the original contributions of this thesis as below:

1. A framework to classify commercially successful villages for underutilised crops through data generated by models developed based on normal crop.
2. A technique to generate realistic and artificial training data using Genetic Algorithm when data for underutilised crops are not available.
3. Automate the labelling of artificial data based on metastable state of fitness value in Genetic Algorithm.

1.7 Organization of Thesis

Structure of this thesis is organized as following:

Chapter 1 introduces the project. Problem statement, research issues, aims and objectives and research scope can also be found in this chapter.

Chapter 2 discusses the importance of commercializing underutilised crops. The chapter also covers past research methods used and current state of generating artificial training samples. Economic Models by other researchers are also

included in this chapter. Contribution of this research to Crops for Future Research Centre is presented at the end of Chapter 2.

Chapter 3 briefly describes the background theories of Genetic Algorithm and Support Vector Machines which aid in understanding development of originality to the methodology.

Chapter 4 presents methodology carried out to fulfil the aim and objectives of this research. The methodology is discussed in detail, which covers constraints setting that reflects real life situations, data labelling method, and also the process and algorithm of generating artificial training data using Genetic Algorithm to train Support Vector Machine.

Chapter 5 contains experimental results on robustness of SVM to handle data without overfitting. Data labelling method described in Chapter 4 is further supported by results that are also presented in this chapter. This is then followed by cross validated results from training SVM using artificial data generated and test results from real village data. Analysis on all experimental results is also discussed in this chapter.

Chapter 6 concludes the research by reviewing the objectives and major accomplishments that are achieved. Potential future work is also included at the end of this chapter.

Chapter 2 Literature Review

2.1 Introduction

The original approach used in this research is a combined machine learning-model prediction strategy in order to provide the capability of using the support vector machine in instances of very low availability of training data.

As such, a method for generating artificial training data (synthetic data) which in this case allows for the high accuracy classification of successful rural villages in cultivating underutilised crops, was developed.

Based on the above, this chapter is organized as follows;

a) A review on the definition and justification of the importance of underutilised crops is given. Discussion of the problems associated with commercializing underutilised crops is also covered.

b) A review of artificial data generation methods used when not enough training data is available. This is relevant in this case because data on underutilised crops which have been successfully cultivated are very limited [3] [15].

c) A review of agricultural economic models from a set of variables which were determined either from surveys and simulation [16], [17]

2.2 Importance of Underutilised Crops

According to studies, 95% of food for the world is provided by only 30 species of plants. 3 major crops which are maize, wheat and rice already made up more than 50% of global consumption [1]. This can be seen from the pie chart as shown where blue, red and green are wheat, rice and maize respectively.

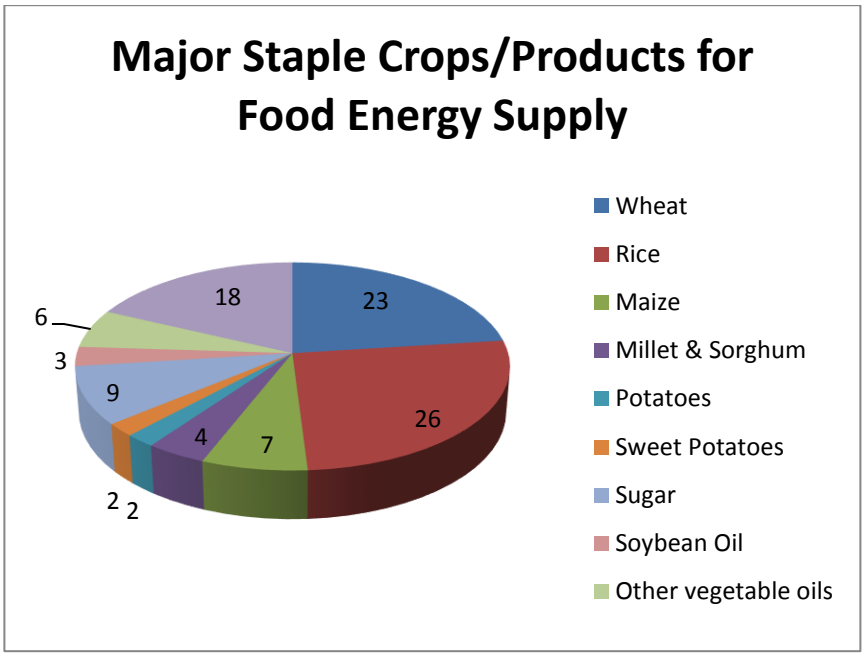


Figure 2.1 Pie Chart of Food Energy Supply for Major Staple Crops

With the economic benefits provided by these major crops, it is believed that these three crops (maize, wheat and rice) will be increasingly dominating the market and consumption.

However, wheat is extremely heat sensitive and rice is vulnerable to heat stress, rising seawater and flooding. Dependence on major crops has inherent agronomic, ecological, nutritional and economic risks that are unsustainable in the long run, especially facing the problem of global climate changes and global warming. It is believed that this will threaten the yield and sustainability of major staple crops directly which, in turn, affects the food security around the world if alternatives are not considered [18].

This is one of the roles that underutilised crops play in ensuring food security. Many of these underutilised species are able to cope with very harsh and difficult environments as some of them are grown wildly in less favourable environments. As mentioned earlier, climate changes and global warming have led us to opt for underutilised crop to ensure food security.

Main concern of this research is to generate income for rural areas. Underutilised crops are normally planted by local people in these areas and used within the community. Commercializing the crops in market helps to raise income of these local people by diversifying plant products. This will contribute to improving the well-being of farmers in rural area.

2.3 Commercialisation of Crops

Commercialisation in general describes the process of making new products available to the market [19]. Commercialisation in terms of farm household is

defined as producing and selling agricultural outputs to the market. This could be done by allocating part of the resources in producing cash commodities which are marketable. Figure 2.2 describes the determinants and effects of commercialisation.

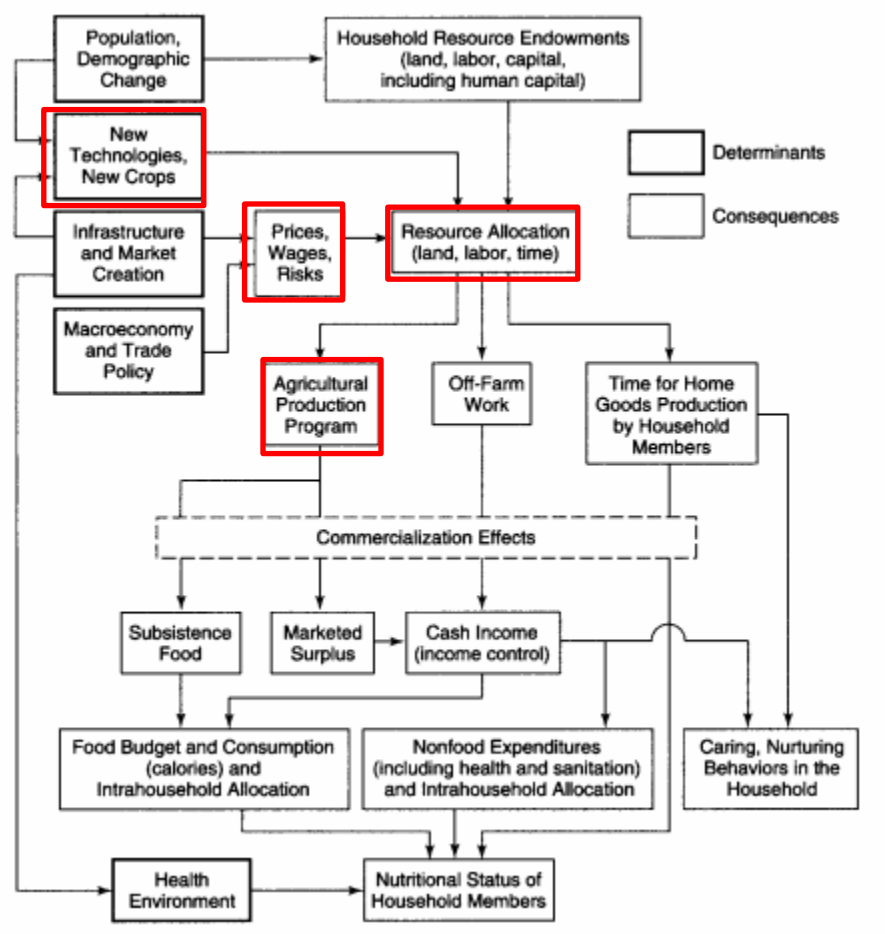


Figure 2.2 Commercialisation Process [20]

Determinants of commercialisation for crops that are identified by von Braun et al. [20] are population and demographic changes, technologies, infrastructure and macro economy policy. In this project the effort to commercialise underutilised crops can be made more efficient by being able to predict the impact of underutilised crops in a certain area even before the crops have been planted. This eliminates the need to plant the crops before deciding on the commercializing prospects, which in turn saves costs and time. With the technique implemented here, the boxes outlined in red in Figure 2.2 and Figure 2.3 will not need to be executed BEFORE the chances of success are computed because the real life data will not need to be collected.

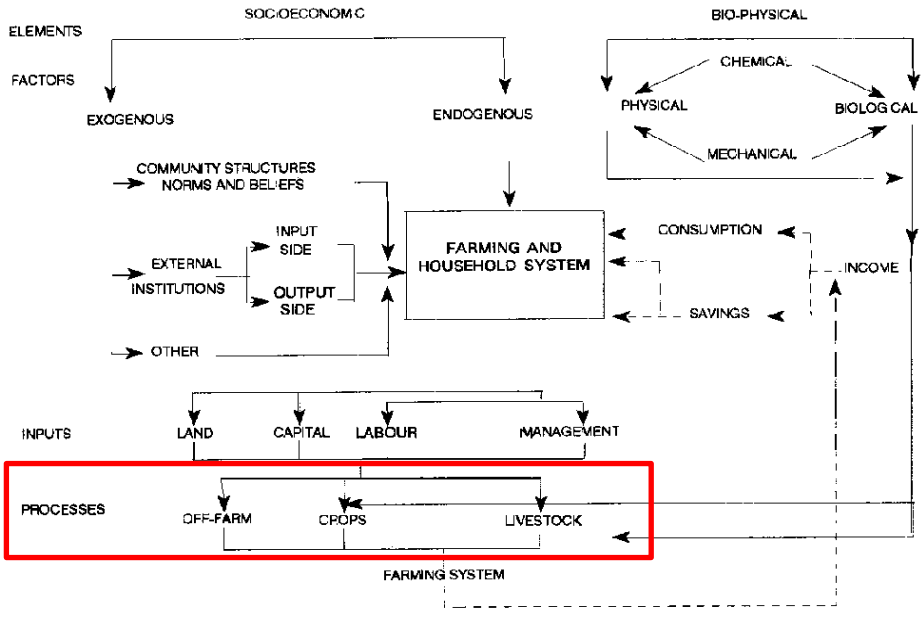


Figure 2.3 Agricultural Commercialisation Process [21]

2.3.1 Promoting Commercialisation of Underutilised Crops

For the past few years, cultivating and trading underutilised crops contributed to food security and food diversification. Selling underutilised crops also has been recognised as a potential income source for poor communities in rural areas in developing countries, as different kinds of policies and framework are being proposed and adapted on underutilised crops to improve their supply and demand. A study carried out by Horna et al. [22] discussed aspects on determining the success of market chains on underutilised crops in Ghana. The crop that is adapted in the paper is garden eggplant in Ghana. The marketing chain of underutilised crops in Ghana is analysed, and a specific policy is suggested to overcome each of the market constraints that are identified for better exploitation of economic potential of the underutilised crop [22].

In another study conducted by Gotor et al. [23], an analysis was done on the impact of interventions of conservation and adaptation of underutilised plant species on livelihoods of people from rural communities. The interventions which include collaboration between farmers and experts in the enhancement of the varieties of the underutilised crops, thereby introducing value-added technologies to the rural community, were implemented in Yemen. Seed Supply System was developed and fertilisation trials were carried out as well. The results

show that these interventions are positively correlated in order to increase the yield of underutilised crop, hence generating more income for the farmers in this rural area [23]. This paper shows that, with proper interventions (such as implementing precision agriculture and value added technologies i.e. broadband technology) to rural farming method could bring more income to local farmers.

The paper discussed by Ravi et al. on mobilizing underutilised crops in India was first initiated from a bench mark survey to study the socio-economic-technological status of the poor community. Based on that, a series of agronomic interventions were suggested to

- promote the cultivation of underutilised crops
- generate income through the use of genetic diversity, improved agronomic practices and marketing [24].

Most of the framework in cultivating and marketing underutilised crops species involved human experts to carry out survey and study on the plants with minimal adaption of technologies. Their marketing strategy was to diversify the varieties of crop products. They aimed to increase the yield of underutilised crops in order to increase income of the rural poor.

In the paper by Guillaume et al. [25], the underutilised crops are defined based on the three characteristics as follows:

1. The crop is locally produced in a certain or restricted rural area;
2. Lack of scientific knowledge on the crop, and only practical knowledge is known by local farmers;
3. The current economic value of these crops is relatively low compared to their potential economic value.

Four main economic factors that classify the underutilised species based on the definition of the paper are [25]:

1. Connection between the current economic value and potential economic value of underutilised crop;
    - The current economic value of underutilised plants can be assessed based on market value and the presence of competitive crops. Local farmers prefer to grow underutilised crops under conditions favourable to themselves; also when commercial crops show a currently high economic value to the local people, which implies that with proper intervention, underutilised crops could be economically successful.
    - The crop and its by-products are only economically valuable if potential consumers have basic knowledge about consuming it. However, practical knowledge about consuming the underutilised crops is only known to local farmers and consumers. Thus, an

information gap is created because of lack of proper channels for transmitting the knowledge to potential users distant from the local market.

2. Absence of an output market for underutilised crops;
    - Output market is said to be “missing” or “absent” when a primary producer is not able to access the market. One of the two reasons for this is high transaction cost which constitutes exogenous constraints causing the local farmers having no access to the market to sell their underutilised products. For example, lack of transport infrastructure can cause this problem.
    - The other factor contributing to the absence of an output market is the cost of handling the underutilised crops. For example, to maintain the freshness on the way to the market and to store them in a proper environment to maintain the shelf life. This factor is limited by endogenous (internal) constraints.
3. Market imperfections of underutilised crops;
    - This can be explained in two ways. Market price of the crop is not the full value of the product or a price which consumers are willing to pay, plus quantity produced is below optimal scale of production. There are some factors that cause these situations, one

being lack of demand due to lack of complete information on the plant product. The major demand comes from local communities of the rural areas where they use the crop in the traditional way passed on by ancestors. The economically competitive cost of major crops also causes the demand to stay with major crops although underutilised crops contain high nutrition value.

- The other factor is the inefficiency of the existing marketing channel for underutilised crops where the process of collecting, distributing, wholesaling, processing and retailing is unorganized. For example the caper buds in Syria is selling at high mark-up price at the end of the supply chain due to lack of transparency and mistrust in middleman. However, at the other end of supply chain, local farmers in rural areas only earn a very small portion of the profit from selling the crops.

4. Market failures of underutilised crops.
    - Lack of incentive for farmers to cultivate underutilised crops because these incentives are provided to produce other major crops instead.
    - Lack of economic information and product knowledge. For example, nutritional benefit provided by underutilised crops is

neglected or unknown to would-be buyers. This lack of knowledge contributes to low demand for the crop product.

Successful commercialisation is defined by two characteristics in this paper [25]. The first characteristic is that distribution of profits should be sufficient for the primary producers to continuously produce the crops. The second characteristic is the sustainability of the market overtime in terms of price and profit margin as demand increases. Based on the economic factors identified, the conditions that affect successful commercialisation of underutilised crops for people in rural area are identified, as follows:

1. Demand expansion;
2. Increase in efficiency of supply;
3. Supply control mechanism.

Figure 2.4 illustrates the requirements of successful commercialisation for underutilised crops. Currently, the economic value of underutilised plant species is low and can be represented by the initial market equilibrium diagram in Figure 2.4(A) as it is produced in small quantities with low prices, plus production is only sufficient for local consumption within the rural areas that cultivate the crops.

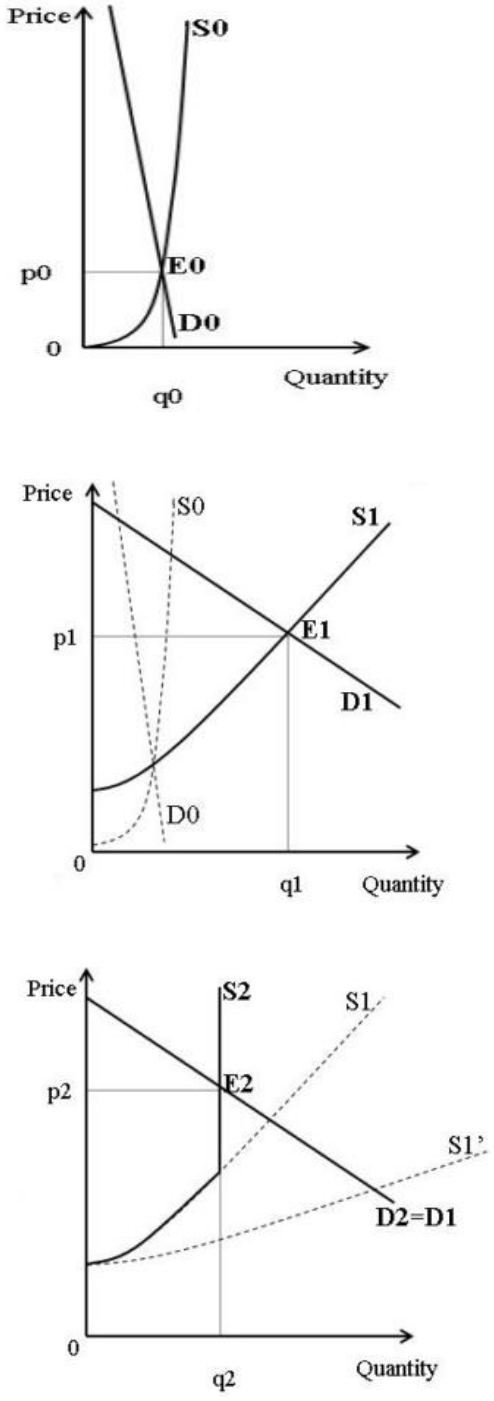


Figure 2.4 Process of Commercialisation Underutilised Crops [25]

Figure 2.4(B) shows the result of expanding demand and increase in efficiency of supply, causing the supply demand curves to move outwards from S0 to S1 and D0 to D1 to new market equilibrium where the price and quantity increase efficiently. This indicates that efficiency of production and marketing system is achieved.

However, increase in the economic value of underutilised crops implies that there can be potential entry of large-scale investments into the market of commoditization mechanism, where increases in efficiency cause the price of products to decrease, thereby influencing the profit margin and incentive of the farmers in rural areas to continue to produce. This can be shown in Figure 2.4(C) where the S1' is the supply curve when the market is commoditized contributing to low selling price, hence low profit margins to the farmers. Thus, supply control mechanism that produced the S2 curve is proposed in this paper. This is to ensure the sustainability of income flow to the poor who grow the crops [25].

2.4 Review of Methods used to generate Artificial Data

Insufficient data is often recognized as the main problem for poor and unreliable performance of machine learning and data mining techniques. This is due to the fact that collecting sufficiently large data for machine learning purposes could be expensive and difficult in some cases [26]. Research on the insufficiency of training data has been carried out in different fields adapting different methods.

Generating artificial data is one of the common practices to generate sufficient data for various research purposes, such as software testing data, machine learning training data, protection for confidential data. Artificial data is also known as virtual samples or synthetic data, and the three terms can be used interchangeably.

2.4.1 Data Generation from distribution function of original samples

Niyogi et al. [27] carried out a study on the creation of virtual examples by incorporating prior knowledge to increase the amount of training data to be used in computer vision and speech recognition problem. Jing et al. [28] generated virtual sample based on Gaussian distribution. This was done by generating random numbers following Gaussian distribution around the original sample data. The approach of using virtual samples generation on evolving fuzzy classifiers was proposed by Hahnel et al. [29] Virtual samples were generated through parametric distribution function where parameters of membership function were mapped to the parameters of the distribution function. This is to ensure that learning from a set of parameters were similar to learning from the original data when virtual samples were created based on parametric class description. The method of generating virtual data from distribution function is by creating values in normalized space and then translating and scaling the virtual samples in class space, which then retransform to original feature space. Narasimhamurthy et al.

[30] proposed a framework to generate artificial data to train and test classifiers in various changing environments, including gradual changes, substitution and systematic trends. New virtual data were sampled from data distribution and added to the current data set to simulate changing environments. Pham et al. [31] demonstrated artificial data generation for intrusion detection. Virtual samples generation for two different situations were proposed: a situation where only normal data was collected with non-existing intrusive data being observed, and a situation where a small amount of intrusive data was collected which are insufficient to train detection model. For fully synthesized data, mean and standard deviation of normal data were calculated. A sample was randomly selected and the values were altered to be out of range of the normal distribution of original data to emulate abnormal behaviour. In the latter situation, samples were randomly selected from the collected intrusive data. Difference between highest frequency of values in identified feature among the samples and frequency of the selected samples was calculated. This was set as the number of artificial data to be generated. Selected samples were altered in within the range of the lowest and greatest values of samples in the original intrusive data as new data. Joaquin et al. [32] generates artificial data sets to evaluate the performance of rule-based classifiers. Four different types of sampling methods were applied to the rules that defined from original knowledge to create artificial samples that

matches these rules. Notake et al. [33] generated synthetic data from elastic wave theory coupled with Gaussian modulated function as time domain to train neural network for crack depth determination. Experimental setup for crack depth in steel plate was configured to verify the synthetic data generated and performance of neural network trained by synthetic data. Zhang et al. [34] proposed an approach of generating synthetic data through parametric model derived from real image data. Synthetic data created and real data were input to multichannel auto encoder (MCAE) to learn from the mapping between these two sets of data. Synthetic data were set as input and real data were set as output for one of the channels, real data were set as input and output for the other channel. With this structure, MCAE is able to learn from two tasks at the same time. By transforming the inputs to the same target outcomes, MCAE minimize the gap between real and synthetic data. Potharst et al. [35] suggested two methods to generate synthetic monotone data based on monotonicity constraint. Figure 2.5 are examples provided to illustrate this method. Figure 2.5 (Top) is the original dataset and translated into Figure 2.5 (Bottom Left). Acyclic model in Figure 2.5 (Bottom Right) is developed. To generate the monotone dataset, all paths in the graph have to be non-increasing where the target label y should not be larger than the previous y. Random features and target labels were selected separately. Target labels were travelled through the path till the end of the nodes until all the paths

were labelled. This resulted in monotone synthetic data. For a more structured data set, acyclic model in Figure 2.5 (Bottom Right) can be replaced by tree based model.

Client	Income	Education	Criminal record	Loan
cl1	Low	Low	Fair	No
cl2	Low	Low	Excellent	Low
cl3	Average	Intermediate	Excellent	Intermediate
cl4	High	Low	Excellent	Intermediate
cl5	High	Intermediate	Excellent	High

x_1	x_2	x_3	y	x	y
0	0	1	0	001	0
0	0	2	1	002	1
1	1	2	2	112	2
2	0	2	2	202	2
2	1	2	3	212	3

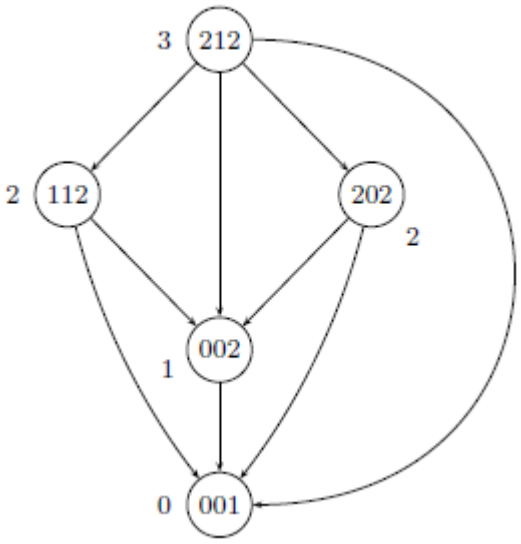


Figure 2.5 (a) Original Data (Top), (b) Transformed Data (Bottom Left), (c) Acyclic model that indicates the path to travel (Bottom Right) [35]

Lin et al. [36] improved the artificial samples for early stage manufacturing process using mega trend diffusion. Principle component analysis was applied to transform the raw data to orthogonal variables to remove dependency between

the variables. Box and whisker plot were used to determine the upper and lower boundaries for triangular membership function. Random values were generated within the boundaries. Patki et al. [37] proposed a synthetic data generation system that they named as Synthetic Data Vault. Users organize data based on structure of the data. The data were organized in such a way that relationships of multiple tables in database are able to be modelled into generative model. Data was tested with Kolmogorov-Smirnov test for the distribution of the data. The generative model was made up by parameters of distribution for each column and covariance matrix. Data were synthesized by sampling from the generative model. This was for fully synthesized data set. To generate partially synthesized data, user was expected to update the covariance matrix and mean vector for the known values.

2.4.2 Data Generation from original data perturbation

Synthetic Minority Over-Sampling Technique is commonly known as SMOTE which is proposed by Chawla et al. [38] to overcome the imbalanced dataset. K nearest neighbours were determined in the minority class. Difference of the sample and nearest neighbour was multiplied by random number between 0-1 to perturb the original sample. This is then added as a new synthetic sample to the minority class. Majority class was also under sampled by randomly removing the sample. On the other hand, Cho et al. [39] created virtual training samples by

selecting points near to the real samples in order to train multilayer perceptron. The approach was demonstrated by generating points that are within a hypersphere with radius of 0.5 centred from real samples. Another approach was also demonstrated by adding Gaussian noise to the real samples. Enzweiler et al. [40] presented an approach using synthesis data to improve pedestrian classification of a hybrid generative-discriminative model. The synthetic data is generated by varying shape, foreground texture and background texture of real pedestrian data. Pishchulin et al. [41] generated synthetic data of 3D shape models by employing rendering-based reshaping method. This was done by varying shape and pose of 3D human bodies. The data generated is combined with original data to train people detection models. Zhan et al. [42] presented virtual samples generation using Neighbour Line based locally linear Embedded to train manifold learning algorithms that required dense samples to perform well. Interpolated points on several neighbour lines from the data were used as virtual samples. Figure 2.6 was used to illustrate the virtual samples generation. Virtual points were interpolated through 6 neighbour lines from data point x_i.

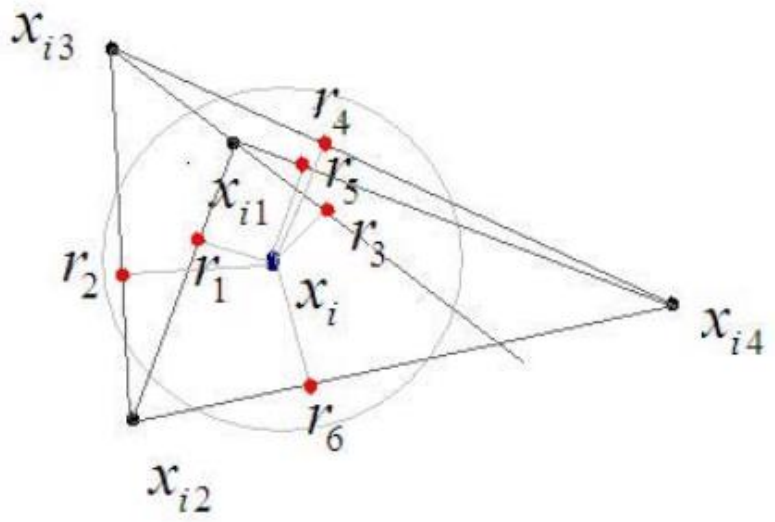


Figure 2.6 Interpolation of neighbour lines from data point x_i [42]

Hu et al. [43] proposed to generate virtual samples through windowed regression oversampling method. The samples and its k nearest neighbours were used to determine regression coefficients. New virtual samples were generated using the regression coefficients through interaction between the nearest neighbour and a pair of randomly selected regression coefficients. Manabu [44] studied the virtual samples generation for text classification to train Support Vector Machine. Virtual samples were created through adding or deleting features in feature vector of the text documents randomly. A fully synthetic handwritten script data approach was proposed by Gaur et al. [45] without any natural handwritten data used to train Support Vector Machines. The synthetic texts were generated from

fonts and distorted to imitate natural handwritten text. Approach proposed by Varga et al. [46] to generate synthetic data for handwriting text recognition was done through perturbation. Instances were generated based on cosine wave function and were distorted through geometrical transformation and thinning/thickening operations. Synthetic data is added to the original data to train the recognition system. Zhang et al. [47] generated virtual face image data by applying random matrices on original face data based on uniform distribution. These virtual samples created were used to validate base classifiers for an ensemble learning framework.

Some researchers proposed synthetic data generator system to create initial set of data based on input parameters from users. Further adjustments were done on the initial data created to generate artificial data. Synthetic data generator system presented by Pei et al. [48] was by applying linear transformation on the initial data set to modify the shapes and densities of the data. Albuquerque et al. [49] proposed adjustment of data structure using three operators. The operators were one-dimensional probability distribution function which controls each of the dimensions individually, two-dimensional PDF which controls two dimensions at the same time, or three-dimensional PDF. In order to imitate the noise in real data, random noise is added to the data as a final step in generating synthetic data. Monekosso et al. [50] proposed a synthetic data generator using perturbation

model on data from environment monitoring sensors for behaviour analysis. A base model was first created based on experimental data from the sensors. The output sensors data were divided into two types, continuous and discrete output. Continuous outputs were converted into discrete outputs and variations were added to the original data to produce synthetic data to model observations from sensors. Hidden Markov Models were used to generate behaviour profile to assess the model. It was found that best performing models were the one with large samples.

2.4.3 Data Generation from other methods

In Haffari et al. [51], a semi-supervised learning and active learning method to build high quality mapping to reduce labelled data for Statistical Machine Translation (SMT) algorithm translation process learning, was proposed. In a study by Lanouette et al. [52], the process used in the pulp and paper industry is optimized by improving accuracy of the model used to represent the system in the presence of limited training data for training a Neural Network. Li et al. [53] on the other hand proposed mega trend diffusion technique and tree based trend diffusion technique to create artificial data samples for countering the problem where the early stage of flexible manufacturing systems produced too little data to train machine learning algorithm for reliable results. Wonders et al. [54] proposed an artificial water flow training data generated by simulating data between two minimum and maximum examples for water and electric meter data.

Research by Jiang et al. [55] adapted a Random Tree that was built from the available data set to generate more virtual data in order to solve the problem of insufficient training data where the number of variables was more than the number of data points. Al Bakary et al. [56] used Genetic Programming Method to construct training data for Back Propagation Neural Network (BPNN) to forecast classes related to the data. The available data was divided into small datasets to represent nodes of a Decision Tree. The Gain Information Ratio of each node was the fitness value of each population. Each population was expanded using Genetic Programming Algorithm before input to a BPNN for classification. Li et al. [57] presented virtual sample generation method through Weibull variate to produce uni-modal and multi-modal virtual samples. Buys et al. [58] illustrated the process of generating human model using open source 3D human illustration known as MakeHuman. Human base mesh was created and pose was applied to the mesh. Accurate virtual measurement was created by placing the human avatar in front of virtual camera in OpenGL environment. These artificial human models were used to train human pose detection machine learning algorithms. In order to train car detector system by providing sufficient car image data from multiple view, Wang et al. [59] generated training samples by collecting data from multi modal sensors on a test-bed vehicle running on the road. Car images collected were pre-processed by removing bad quality images

before training the car detector.

Data synthesis was also extended to be applied on multi label datasets [60][61]. In the framework developed by Tomas et al. [60], users were required to input the parameters to generate the synthesized data. The synthesized data were based on two strategies where users can choose which to apply, HyperSpheres and Hypercubes. In HyperSpheres, a hypersphere was created centred on the origin of Cartesian coordinates system. Small hyperspheres were generated within the hypersphere created. Random points were generated within these hyperspheres with multi labels included. Implementation of HyperCubes strategy was similar to HyperSpheres, except that it was more suitable to be used to evaluate classification algorithms that use hyperplanes. For the framework presented by Read et al. [61], single label virtual samples were generated. Feature space and label space of the single label virtual samples were combined to transform into multi label virtual samples.

Artificial data generation is not limited only to encounter situations where training data is insufficient. Artificial data plays an important role to protect confidential data by extracting general properties of the confidential data and generating artificial data based on these properties [62]. Drechsler [63] proposed to use Support Vector Machines to generate artificial data for sensitive data protection. Joint distribution of original data was modelled using SVM. Sensitive

and private data were imputed partially with the synthetic data generated from SVM. Data utility and risk disclosure were evaluated to validate the synthetic data generated by SVM. Lee et al. [64] investigated the application of Classification and Regression Tree on confidential business data. The values of sensitive variable to be synthesized were the outcome and the rest of the data set was the predictors. A tree will be fitted for outcome and predictors. Synthetic values were generated by following the split up to the terminal leaf and then drawing them with selection probabilities generated by Bayesian bootstrap. Evaluations of the partially synthesized data were made based on disclosure risk and data utility. Comparisons with traditional statistical disclosure control and remote analysis method were also carried out.

2.4.4 Analysis on Methods of Artificial Data Generation

Genetic Algorithm is frequently being used to optimize parameters for SVM [65]–[67]. However, the idea of generating training data for machine learning purpose using Genetic Algorithm is not common. Li et al. [68] proposed to generate virtual sample by implementing Genetic Algorithm, in two main steps: feasibility based programming model is constructed using backpropagation neural network, Genetic Algorithm is implemented to accelerate feasible virtual sample generation. However, the implementation is significantly different. One obvious difference is that their proposed method is used for regression and our framework is on classification. In their method, small data sample are required

in order to use backpropagation neural network (BPNN) to build feasibility based programming (FBP) model. A Megatrend Diffusion Function is used to determine the ranges of each attribute. Difference between the predicted value and the real value by FBP modelling is used as the fitness function in GA. Every population generated by the GA is evaluated by calculating the mean absolute percentage error (MAPE). This sample population is retained if the calculated MAPE is within acceptable ranges. These steps are repeated according to the number of virtual samples that needs to be generated. In our framework, using the GA to generate synthetic data is more effective due to the natural way that labelling of the data can be achieved based on schema theory. The GA generated synthetic data to train Support Vector Machine for classification where the constraints and boundaries would be set to reflect real life situations. With the constraints reflecting real life situations, the training data generated is an approximate representation of the real data. Details of the methodology will be discussed in Chapter 4.

Methods proposed by the researchers that has been discussed were well suited for their areas of applications where small amount of historical data is still available. In this project, there is completely no historical data for the domain in question. The reason for complete lack of data in the commercialisation of underutilised crops is due to nonexistence of any large scale implementation

related to these crops. As such a new method for generating the relevant data was required. Artificial data generation is proposed in this project to overcome absence of relevant data through approximate models.

Three criteria are needed to be taken into account to evaluate the methods of artificial data generation. First criterion is the proximity of the artificially generated data from the proposed methods to the real data. Second criterion is how feasible are the proposed methods to be used on different applications. Third criterion is complexity of the methods. Balancing between all three criteria for a synthetic data generation method is not always achievable. Data generated from distribution function of original samples [27]–[36] usually fulfil the first criteria, while data generated by perturbation of original data [39]–[42], [44]–[47] fulfils the second criterion. General frameworks for generation of synthetic data [37], [48]–[50] are designed to be used on different types of data, fulfilling the third criterion.

This framework is adaptable to any applications without data but where approximate models can be derived. The methodology of this framework is straightforward which makes it simple to be implemented. Synthetic data generation method proposed in this project is fulfilling all three criteria.

2.5 Data Labelling

Data labelling is important task in supervised learning in order to ensure the supervised learning algorithm is classifying properly by feeding the machine with good quality data. Good quality data should protected from being mislabelled and label being missed out from data [69]. IBM is hiring people to only focus on data labelling manually due to the importance of labelled data as emphasized by head of team who are creating Watson in IBM. Data labelling is important for machines to be able to learn even for unsupervised data as stated by Professor of AI at the University of New South Wales [70]. Self-labelled techniques mainly focused on using semi-supervised learning algorithms on partially labelled data [71]–[74]. In this research, the GA generated data is labelled based on fitness value generated from GA itself. The details will be described in the next chapter of this thesis.

2.6 Economic Models

Economic models are mathematical descriptions in the form of equations that contain significant variables which impact the economic status of a certain group of people. Economic models are formulated by economists from collected data, economic theory and targeted questionnaires in order to deduce future results from past occurrence. Generally, there are three types of economic models: theoretical models, mathematical models and simulation models. Mathematically approached economic models are derived to represent economic behaviour under

the assumptions of a specific objective being maximized subject to a set of predefined constraints. The models served to answer to economic phenomena qualitatively. Empirical models on the other hand verify qualitative predictions from theoretical models through data collection for each variable in the model. These models provide estimated numerical value by applying statistical techniques on the data. Simulation models are made up by mathematical models with the combination of logical concepts to reflect real life economic situation. Simulation models start with a set of initial values. These values may change or remain constant during simulation and depend on the conditions set by users. These models are able to show the secondary effect and hidden feedback which theoretical models and empirical models are not capable of [75], [76].

2.6.1 Economic Models based on Mathematical Approach

Barnum et al. [77] presented an econometric model for a semi-commercial family farm to study the integration of household production and consumption decisions on farm profit. Demand function and profit function were derived from Cobb-Douglas Production function. Consumption function was derived from linear expenditure system. Husin [78] demonstrated analysis on factors affecting farm household economic decisions in order to secure food within the household. Economic behaviour is derived from utility maximization function. It was found that in order to increase purchasing power and food security in the farm household, the farm families were advised not to rely only on one main source of

income. Bekele et al. [79] established relationship between agricultural commercialisation and crop productivity of smallholder farms in Ethiopia. This was done by measuring crop productivity and commercial orientation in simultaneous regression model framework. Pender et al. [80] developed a theoretical model based on determinants influencing farm household crops commercialisation. Hypotheses that were deduced from the theoretical model were then tested with data collected from villages in Ethiopia. In a paper published by Niragira et al. [81] was discussed how crop production is affected by agricultural specialization through mathematical programming farm model. Farm households are categorized into large, medium, small and landless farms. Market values of crop production were maximized as objective function of the model, subject to constraints on resources that were available. Analysis was made based on the output from this model. In a study by Groothuis [82], mathematical programing model was employed to examine the relationship between farm income and production of perennial crops. Farm household income was formulated as an objective function to be maximized subject to resources constraints. Eaton examined the relationship between farm productivity and soil erosion. Mathematical equation for farm income was rearranged to determine economic productive life of soil. Results were analysed by observing the change on economic productive life of soil when parameters in the equation were varied.

Mathematical models are used to describe simplified complex real life economical situations. This aids in understanding an economic phenomenon. However, due to the simplicity of a model, it is questionable that the model is able to capture key elements of a real world economic situation. The model may only works in describing certain situations under certain assumptions.

2.6.2 Economic Models based on Surveys (Empirical Approach)

Okezie et al. [16] studies the factors that drive agricultural commercialisation in Nigeria. It is found that using Cobb Douglas production function, the determinants of agricultural commercialisation were labour, fertilizer and planting materials Adong et al. [83] did a survey in Uganda to model market participation seasonally in commercialisation. It is found that the factors that affect market participation are varied depending on the season of the year; also the locations throughout the country. In India, Rahut et al. [84] collected household data through questionnaires to explore determinants that affected agricultural commercialisation. Land size, gender of household head, livestock assets, ethnicity, education and location were identified as significant factors that affect commercialisation of agriculture. Data based on questionnaire were collected by Aderemi et al. [85] in Nigeria to investigate the farmers' characteristics that determine agricultural commercialisation. Among the twelve factors that were identified, it is found that information on price and commodity, household size, farm size, farming experience and access to extension workers

were significantly influencing household commercialisation. A household survey was carried out in Ghana by Samuel et al. [86] to analyse the relationships between commercialisation of tomato and pineapple in farm and the factors that determine commercialisation of small holder farmers. The study also discusses the impacts of commercialisation on food security among the farmers. In order to study the factors of commercialisation, questionnaires were used to collect data in Abia State of Nigeria by Oteh et al.[87] It was found that six out of ten of the variables from the multiple regression models were important factors influencing commercialisation in Abia. These six variables were value of output, farm size, sex, distance to market, membership of cooperative and farming experience. Seyoum et al. [88] conducted a survey in Kombolcha district of Ethiopia to examine degree of commercialisation for potatoes. Farm size, access to irrigation as well as to market information were among 14 variables that made up the economic model, and determined as major determinants affecting commercialisation of potatoes. Another data collection based on questionnaire was carried out in Abia, Nigeria by Agwu et al. [89] to analyse factors of agricultural commercialisation among small holder farmers. Multiple regression models were formed based on ten variables that were identified. Household size, income, farming experience, farm size, distance to market, membership of

society and access to credits were found to be significant determinants which affect commercialisation in the area.

As can be seen, the variables that form the economic models are similar but different although the models were all examining agricultural commercialisation. A few of the models were for the same area. These economic models were formed largely based on authors' assumptions. Much data is also needed to develop a valid empirical model.

2.6.3 Economic Models based on Simulation

In order to study a more complex system on crop farming, simulation models are used as there are more variables and parameters involved in such models. Bobojonov et al. [17] studied on impacts of climate change on farm income by using bio-economic model. The bio-economic model took agro-ecological and socio-economic characteristics of Kazakhstan, Kyrgyzstan, Tajikistan and Uzbekistan as input variables. Yields of the crops (cotton, potatoes and wheat) were observed, and analysis on farm income according to the crops yield was discussed. In a paper by Kikuhara et al. [90], a bio-economic model was developed to study dairy and crop production in Japan. The model was used to observe the impacts of changing economic and environmental variables on farm output and profitability. Kruseman et al. [91] presented a bio-economic model to evaluate the effects of management decisions on farm household income and

soil fertility. The model incorporated production activities module, expenditures module, prices module, resource endowments based on farm household stratification module, savings and investments module and objectives and goal weights module. Quaranta et al. [92] proposed a bio-economic model that took into account price system, farm characteristics, socio economic feature of family, and physical environment to simulate farmer behaviour that affects the farm management decision in soil risky area of Mediterranean. Mathematical programming was used to simulate and calculate farm household decisions and practices which affect family income. Econometric process model and bio-physical model were integrated by Antle et al. [93] to simulate farmers' decision making process based on soil, climate and economic conditions. The model was able to simulate net returns, distributions and supply functions for crops within the samples, and predict 'out of samples' distribution.

Although simulation models require less data as compared to empirical models, sufficient data is still demanded in order to develop mathematical functions which are the core for simulation models. Keeping the data up to date is also crucial especially as it will affect parameters in the model. It is not possible however for certain natural disasters (such as earth quake) or certain economic conditions (such as economy downturn) that could affect the farm production are

to be represented in mathematical function. As a result, outputs from the models may not be accurate.

2.6.4 Other Economic Models

2.6.4.1 Economic models related to rural economic growth

As mentioned in previous chapter, various studies on economic models were carried out for different purposes. Shang et al. [94] forecasted rural GDP by using Grey Model. Grey Model uses first order differential equation that indicates exponentially increasing trend. However, relationship between rural GDP and factors is lacking from this type of model. Issam [95] has adapted system dynamics models to analyse macroeconomic factors in Yemen. The model takes into account a wide range of macroeconomic variables that relate to GDP. The variables included are population, crops production, fishing production, oil and gas production, and manufacturing production. Analysis by this study is too broad, which makes it unsuitable for use in this research. Relationship between GDP and output of selected major cash crops in Nigeria is examined in a paper by Anyanwu et al. [96]. This is not suitable to be adapted as the types of crop that affect national GDP are not the concern in this research. In Nigeria, Eravwoke et al. [97] studied major cash crops (oil palm, cotton, rubber, cocoa and groundnut) production and transport on economy of their country.

2.6.4.2 Economic models related to cash crops

A model that captures supply and demand of wheat crops in Pakistan is used by Zulfiqar et al [98] for policy analysis. This model examined supply and demand at national levels. It does not take into account the village characteristics impact on wheat crops farming. Hence, it is not suitable for use in this research. Gerald et al. [99] presented the effects of climate change on crops by incorporating climate, crop and economic models. Paloma et al. [100] has presented integration of an economic model and a hydrologic model where the outputs from one model were used as inputs for the other one. This framework was used to evaluate impacts of climate change on crops and irrigation system. Hans [101] did a study on the effect of cereal aphid on crop yields through crop loss model. A dynamic economic threshold was suggested through the analysis of the proposed model. Crop yield prediction models were presented by Askar et al. [102] by applying the models in different districts of Ghana. Garcia-Vila et al. [103] uses AquaCrop model to evaluate crop yield in response to irrigation provided to the farm. Outputs from the model are used as inputs to economic optimization model to optimize irrigation system when water supply to the farm is scarce.

2.6.5 Limitations and Inaccuracies of Economic Models

Formulating economic models boils down to two critical elements which are data and mathematical functions. These elements are also main causes of the limitations in economic models. Human behaviours and real world systems are

too complicated to be described using mathematical equations. It is also impossible to take into account all important aspects in real world system in modelling. Significant variables could be inadvertently omitted from the models which lead to their inaccuracy and infeasibility. Mathematical models were formulated based on data. However, economic data could be unavailable or incomplete. The randomness of data causes inaccuracy of the output from the economic models. Assumptions are made when economic models are created. The assumptions made are usually subject to the economists' judgements on their interpretations of real life. As observed from the empirical models in Section 2.6.2, for the same subjects that were to be examined, different variables were involved in the models. Historical data could be outdated when new economic policies or adjustments on existing policies are introduced in the country. Using these outdated data as foundation to formulate economic models induce inaccuracy in the model which then produces misleading outputs. Models calibration is also subject to inaccuracy. An interesting experiment was carried out by Jonathan Carter on geophysical models. He claimed that a model describing the conditions of a hypothetical oil field that was set up by him was a perfect model. A set of arbitrary parameters was used on the model to produce data. Data produced from the perfect model were perfect data. The parameters were then removed for perfect data to calibrate the model, expecting the same set

of parameters to be produced. However, it was found that the same set of data could be produced from a different set of parameters for the same model. This is the result tweaking the value of terms and parameters in various ways in the mathematical equations that could add up to the same output [104]. Hence, different models could fit into the same historical data which causes the inaccuracy of output from the inaccurate models. In methodology that is proposed by this research, this shortcoming is overcome by combining modelling with machine learning. The machine learning improves the accuracy of the system by assigning class instead of calculating a single solution in the search space.

2.6.6 Machine Learning vs Economic Modelling

In machine learning and economic modelling, data are equally important to build and validate the model. Ultimately, both machine learning and economic modelling are used to achieve the same goal, which is to learn from data for future predictions. Machine learning is able to deal with a wide range of features and a big number of instances. Without the prior knowledge on underlying relationships between the attributes, machine learning is able to learn and discover patterns within the data and make predictions for new data. In contrast, economical modelling requires modellers to understand relationships between variables before modelling. Economical modelling is based on various sets of assumptions when collecting data, formulating the model and analysing the

outputs. For examples, in empirical models, dependent and independent variables are assumed to be linearly related. Implicit assumptions that the dependent variables affect independent variables are usually made before data is tested. Some other general assumptions for economic models are that all people are assumed to be rational and all people want to achieve utility maximization [105] [106]. Figure 2.7 shows an interesting case study [107] from McKinsey & Company on evaluating the risk level of customer leaving the service of a telecom company. Green line in the figure indicates the classic regression analysis which

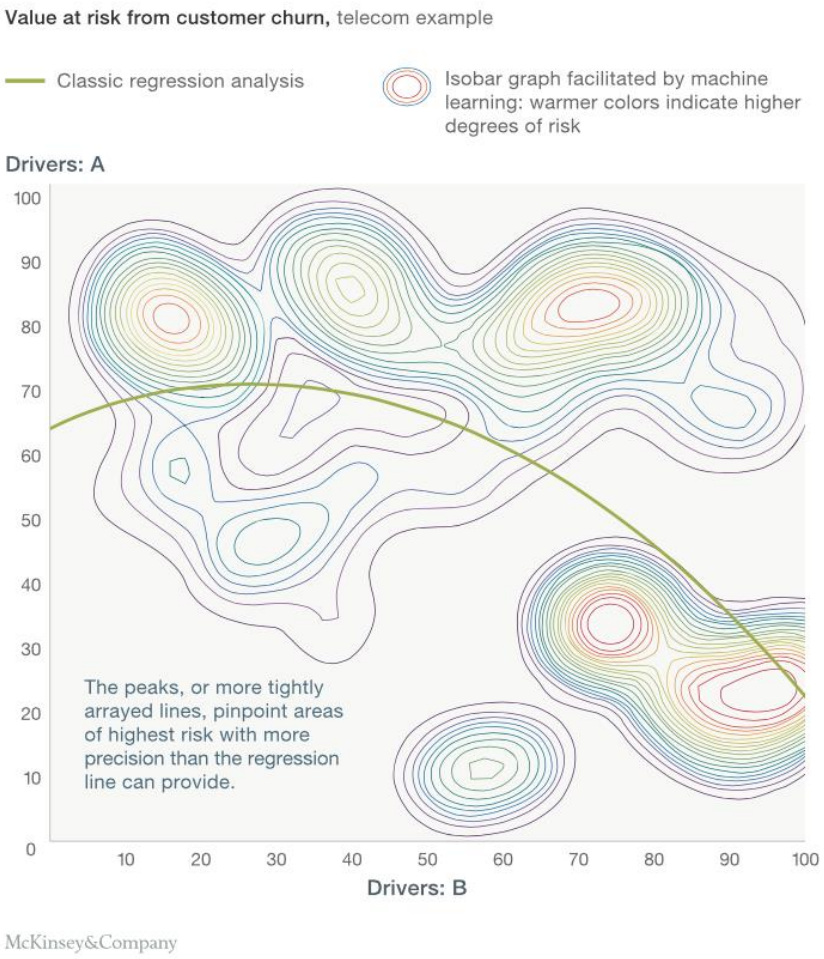


Figure 2.7 Case Study on evaluating the risk level of customer leaving the service of a telecom company [107]

is the technique used in formulating empirical models. This figure clearly shows what machine learning can do which empirical models cannot. Machine learning is able to capture patterns of the data beyond linear boundaries, whereas, empirical models can only formulate simple regression line for the data [108]. Nonetheless, economic models focus more on explaining and reasoning

economic phenomena than classifying and predicting as to whether machine learning is obviously performing far better than economic models. In Chapter 5, experiments are used to describe how economic models and machine learning complement each other to overcome limitations of a model and solve classification problems.

2.6.7 Review on Methods of Improving Traditional Modelling Methods

As addressed earlier with regard to the issue of limitations and inaccuracy of traditional models, some existing research works to overcome these limitations will be presented in this section. Generally there are two ways to handle the limitations of traditional modelling methods, i.e. by substituting machine learning models in place of traditional models and by integrating machine learning models with traditional models to form a robust hybrid model.

2.6.8 Substitution of Machine Learning Models to Traditional Models

Laura [109] demonstrated the superiority of Artificial Neural Networks over classical discriminative analysis on building a classification model to target market strategies to the identified group of consumers. Shahrabi et al. [110] compared the performance of machine learning techniques and traditional forecasting models. The research found that machine learning techniques outperformed traditional models in forecasting supply/demand chain of car components. Machine Learning techniques can also successfully overcome prediction on economic turning point [111][112] which traditional economic

models usually fail to forecast [113] due to randomness and noisiness of economic data.

2.6.9 Integration of Machine Learning Models to Traditional Models

Mora-Lopez et al. [114] presented a hybrid model between multivariate regression model, artificial neural network and probabilistic finite automata on forecasting atmospheric clarity value. The study shows that hybrid models between traditional model and machine learning model yield better forecasting results as compared to the individual model. Xiang et al. [115] developed a hybrid model of autoregressive integrate moving average (ARIMA) in conjunction with Support Vector Regression to predict turbo generator vibration by taking advantage of superiority of ARIMA on linear pattern and SVR on nonlinear pattern.

2.6.10 Analysis on Methods of Improving on Traditional Modelling Methods

While there are many methods proposed by researchers to improve the limitations and accuracy of traditional forecasting models, these models require historical data in order to carry out prediction. Using these methods does not solve problems with limited historical data when new policies are introduced or existing policies is adjusted. Hence, the framework proposed in this project is well suited to address these two main issues that are discussed earlier. The proposed framework generates sufficient data using Genetic Algorithm from econometric model which approximate real data in order to train SVM. In

addition, these economic models are developed for normal crops based on village characteristics data collected for normal crops. This has motivated the research to focus on building a framework for underutilised crops to help farmers in determining whether or not to cultivate new underutilised crops. The proposed framework predicts the potential as to whether underutilised crops will be successful in bringing economic growth to farmers based on the village characteristics data by using a combined machine learning and modelling method. This is done before having the crops planted physically to save a tremendous amount of cost and time.

Chapter 3 Relevance of Support Vector Machine and Genetic Algorithm

3.1 Introduction

In this chapter, Schema Theorem of Genetic Algorithm is discussed. This explains how data labelling in this framework is feasible. Background theories of SVM are presented to justify the proposed framework.

3.2 Genetic Algorithm

Genetic Algorithm (GA) is adaptive heuristic search algorithm with its numerical optimization inspired by both natural selection and natural genetics [116] invented by John Holland in 1975. GA is implemented by initializing a random population of chromosomes and going through three operators, i.e. selection, crossover and mutation, to obtain the fittest solution for a wide range of problems. The main idea of GA as part of evolutionary algorithm is based on Darwin's theory about evolution, the survival of the fittest concept. Each feasible solution is encoded as chromosome.

The terms that are commonly used in GA are as following:

- Chromosome / Individual strings – a set of genes that contains the solution.
- Gene – a part of chromosome that determines the solution.
- Population – a set of solutions with the same length of chromosome.

- Elitism – a process where the best chromosome found so far is copied to populate the next generation before crossover and mutation are carried out to preserve the best solution. By using this method, performance of GA is greatly increased as the population converges quicker than usual. [117]
- Fitness function – function that assigns fitness value to each chromosome.

The fitness of each chromosome is determined by fitness function, and this in turn determines ability of the chromosome to reproduce offspring for next generation. After each iteration, the solutions are given a performance measure derived from fitness function, and the best solution of the population will then propagate to next generation. This made the fitness function as the most important element in GA. The fitness of a solution depicts the closeness of the solution to the optimal value.

3.2.1 Schema Theorem

Schema is a template that contains similar subset of strings of '0','1' and '*' where '*' is wildcard that can represents '0' or '1'. For example, [0 1 1], [0 1 0], [0 0 1] is identified by schema [0 * *]. Each gene in Genetic Algorithm is represented by a string of such bits. Any specified chromosome is an instance of many schemata. Hence, each time when evaluating fitness of a given chromosome, information about average fitness of each schemata is gathered.

Order of schema $o(H)$ is the number of non-wildcard in the schema. Defining length $\delta(H)$ is defined as the distance between first and last non-wildcard gene in the schema. Hence, $o(0**)$ is 1 for the one non-wildcard gene and $\delta(0**)$ is 0 because the distance between first and non-wildcard is 0. Holland proved that by applying genetic operators such as crossover and mutation, each schema represented in the current population will increase or decrease in line with its relative fitness, independent of what happens to other schema. This is known as schema theory. If crossover and mutation is involved, instances of schema H may become some other schema H'. This process is described as destruction. While at the same time, in some other instance H' may become schema H through recombination and mutation. This process is described as construction. These processes are neglected in order to state the result about schema H independently of all other schemata.

The probability of selecting a particular schema H is:

$$\frac{f(x_i)}{\sum_j f(x_j)} \tag{3.1}$$

Every individual in population n is selected individually. The number of selected individuals that fulfill schema H is binomially distributed with sample amount n and probability above. Hence, expected number of selecting an individual which fulfil schema H is:

$$n \cdot \frac{f(x_i)}{\sum_j f(x_j)} = n \cdot \frac{m(H,t)}{m(H,t)} \cdot \frac{f(H)}{\sum_j f(x_j)}$$

$$= m(H,t) \cdot \frac{\frac{f(H)}{m(H,t)}}{\frac{\sum_j f(x_j)}{n}}$$

$$= m(H,t) \cdot \frac{\bar{f}(H,t)}{\bar{f}(t)} \tag{3.2}$$

$\frac{\bar{f}(H,t)}{\bar{f}(t)}$ is the selection term from fitness proportional selection. For individuals which matches H are above average fitness, they will have more children than the rest of the population. Hence, lower bound for $E[m(H,t+1)]$ increases or decreases in proportion to their fitness divided by the average fitness.

Probability of schema H destroyed by crossover is:

$$p = \frac{\delta(H)}{l-1} \tag{3.3}$$

This means that it is a chance of choosing a crossover point that lies between defining length of H.

The probability where a gene that fulfils H produces offspring that also fulfil H is survival probability and defined as:

$$p_s \geq 1 - p_c \cdot p_d \tag{3.4}$$

This crossover term represents lower bound for the probability that the schema H will survive crossover at generation t. This is expressed as one minus upper bound for the disruption probability. The upper bound is represented by $p_c \cdot p_d$ where p_c represents the probability of performing crossover as opposed to selection followed by cloning. p_d is made up of two terms. The first term is probability of schema H destroyed by crossover. Second term places an upper bound on disruption by calculating the chance that the other parent also matches schema H and subtracting it from 1 to give the chance that the second parent does not match H.

$$p_d = \frac{\delta(H)}{l-1} \cdot (1 - \frac{m(H,t)f(H,t)}{M\bar{f}(t)}) \tag{3.5}$$

Probability of H survived mutation is

$$(1-p_m)^{O(H)} \tag{3.6}$$

This represents the mutation term where the proportion of children that match Schema H which still match H after mutation.

By combining equation (3.2), (3.4) and (3.6), Schema Theorem is expressed as:

$$E[m(H,t+1)] \geq m(H,t) \cdot \underbrace{\frac{f(H,t)}{\bar{f}(t)}}_{Selection} \cdot \underbrace{(1-p_m)^{O(H)}}_{Mutation} \tag{3.7}$$

$$\cdot \underbrace{\{1 - p_c \cdot p_d(H,t)\}}_{Crossover}$$

Where $m(H,t)$ is the number of strings matching the schema H at generation t. $f(H,t)$ is mean fitness of strings matching H. $\bar{f}(t)$ is the mean fitness of the strings in population. p_m is the probability of permutation per bit. p_c is the probability of crossover. l is the number of bits in the strings and $E[m(H,t+1)]$ is expected number of strings matching the schema H at generation t+1.

From Equation 3.7, it can be seen that the fitness of solutions in each generation is affected by the two genetic operators, selection and crossover. Selection is carried out by evaluating the fitness of the strings at current generation based on fitness function and selecting the strings based on better fitness value. The parent strings that are selected perform crossover to produce better offspring. Mutation is important to ensure genetic diversity within the population in order to avoid premature convergence to suboptimal solution. Premature convergence occurs when the strings became too similar to one another as an effect of selecting the fittest strings in the population for each generation. Referring back to the same expression, it can be concluded that short, low order and above average schemata (building blocks) increase exponentially in subsequent generations as they combine with each other to form better solutions. The way of how GA works is

named as building blocks hypothesis by Goldberg [118]. Building blocks hypothesis described the process of convergence within GA by combining schemata with short, low-order and fitness above average to form higher order schemata over and over again until the convergence of optimal solutions.

While the inequality (Equation 3.7) is formed for genewise mutation, one point crossover and fitness proportional selection, a more general form of Schema Theorem is:

$$E[m(H,t+1)] \geq \{1-\eta(H,t)\}\,\mu(H,t)m(H,t) \qquad (3.8)$$

where $\eta(\cdot)$ represents loss term relating to transcription error effects of recombination and mutation while $\mu(\cdot)$ is selection coefficient. By stating Schema Theorem in this general form makes it more obvious to Holland's basic idea that there are two competing forces in GA, one which increases the representation of good schemata in line with their fitness while the other tends to break them up.

3.2.2 Modification from Schema Theorem

Schema Theorem is criticized as the effect of creation by crossover and mutation is neglected by Holland. Hence, Schema Theorem is expressed as lower bound instead of an equation on expected number of offspring in the next generation which match schema H. Stephens and Waelbroeck [119] extended the Schema Theory for GA to exact Schema Theory by partitioning schemata into 'left' and

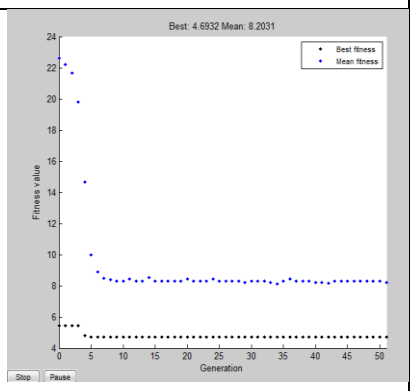
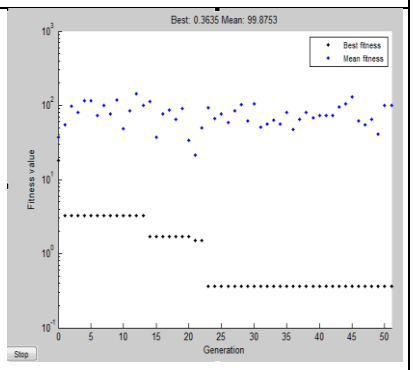
'right' pairs of shorter schemata than the original schemata. In population of size n, $E[m(H,t+1)] = n\alpha(H,t)$ where

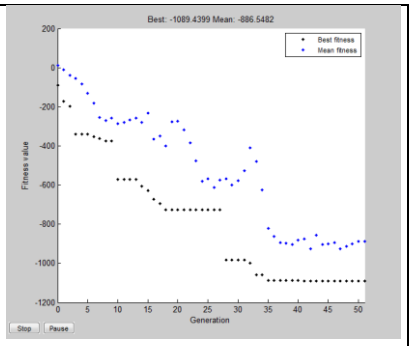
$$\alpha(H,t) = (1-p_c)p(H,t) + \frac{p_c}{l-1}\sum_{i=1}^{l-1} p(L(H,i),t)p(R(H,i),t) \tag{3.9}$$

where $L(H,i)$ is the left partition of schema H, which the schema obtained by replacing 'don't care' symbols of all elements in H from position $i+1$ to position l. $R(H,i)$ is the right partition of schema H, which the schema obtained by replacing 'don't care' symbols of all elements in H from position 1 to position i. $p(H,t)$ is the probability of selecting an individual matching schema H to be parent.

In a study to examine mutation rate on GA, Vose [120] found that change of value on mutation rate caused great change in GA's trajectory which is not obvious enough to be explained by Schema Theorem. A dynamical system model takes into account both construction and destruction of schemata by considering all possible ways of performing crossover and mutation. The model provided an insight on the expected behaviour of generic GA that involves selection, crossover and mutation for finite population. The population is expected to move quickly to a metastable state close to a fixed point and stay there for some time before escaping into another metastable region depending on the fitness function.

Table 3.1 Examples on metastable state of different fitness function

Test Function Name	**Fitness Function**	**Generation VS Fitness Value**
Ackley's Function	$f(x) = -a\exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d}x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(cx_i)\right) + a + \exp(1)$	
Rastrigin Function	$f(x) = 10d + \sum_{i=1}^{d}[x_i^2 - 10\cos(2\pi x_i)]$	

Egg Holder Function	$f(x) = -(x_2 + 47)\sin\left(\sqrt{\left\lvert x_2 + \frac{x_1}{2} + 47\right\rvert}\right) - x_1 \sin(\sqrt{\lvert x_1 - (x_2 + 47)\rvert})$	

quickly to a metastable state close to a fixed point and stay there for some time before escaping into another metastable region depending on the fitness function. This can be observed in Table 3.1 which shows some examples of metastable state of different fitness function.

Metastable state is common in a dynamical system where the system experiences a stable state for a long period of time before a rapid change occurs. Nimwegen et al. [121] introduced a mechanism that predicts the occurrence of the stable state to understand the metastability behaviour in Genetic Algorithm. The study found that the change in metastable level of fitness (which is known as 'fitness epoch' in the paper) for GA that includes crossover is the same as GA with only mutation operator. However, at the beginning, the number of initial epochs is smaller with crossover than that without crossover. It is then deduced that the population will not converge because crossover combines different building blocks to achieve strings with better fitness. After crossover puts together the best

string that can be produced in initial population, the population converges to the best string. During this stage, the building blocks are mostly disrupted by mutation, which explains how the epoch level is at the same fitness level without crossover. The sudden transition from metastable state to another equilibrium state after a long time is known as punctuated equilibrium [122]. Punctuated equilibrium describes the species evolution process where huge changes caused by mutation in the genes of a few individuals are helpful in the survival to the environment, and then settling down to a period of little change. This can be observed from fossil changes and as illustrated in Figure 3.1.

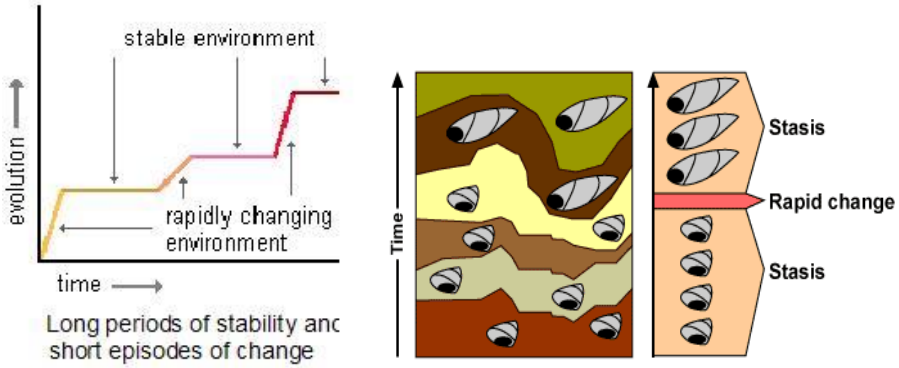


Figure 3.1 Punctuated Equilibrium of fossils [123]

3.2.3 Section Summary

As such the metastable state is utilised to automatically label training data in this research into two classes, one before the epoch and the other class after the epoch. The set of data which is converging to optimal solution is labelled as potentially

successful class. The other data is labelled as potentially unsuccessful class. This labelled data is used to train SVM to classify the economic potential based on new village data on underutilised crops.

3.3 Statistical Learning Theory

Process of learning from examples can be described in three steps:

- A set of examples is collected.
- A model is constructed by learning from the samples.
- Implementation of the model on unseen samples.

Given a set of data $(X, Y) = \{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$ which is randomly drawn from unknown distribution of $P(X, Y) = P(X)P(Y|X)$, the goal is to learn target function $t(X) \rightarrow Y$ that will correctly classify unseen samples Y given X. For binary classification learning problem, Y is restricted to output of two classes where $Y = \{-1,1\}$. Performance of classification can be measured using the risk of learning machine misclassifying training data that evaluates target output and actual output produced from the hypothesis, $f(X)$ and this can be expressed as a function called the "risk functional". The real risk is defined as the expectation value of loss between output given by hypothesis function, $f(X)$ and actual output Y in joint distribution $P(X, Y)$ [124]:

$$R(f) = \int L(f(X), Y) dP(X, Y) \tag{3.10}$$

where $L(f(x), y)$ is the loss function.

3.3.1 Empirical Risk Minimization

Evaluation of the real risk cannot be measured directly since the joint distribution $P(X, Y)$ is unknown. Instead, empirical risk is used by averaging the error function on the training set.

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} L(f(x_i), y_i) \tag{3.11}$$

The concept of ERM is to find the function that minimizes average risk on training set, provided there is sufficient training data available because law of large numbers ensures that empirical risk will converge to real risk when the number of data approaches infinity. Law of large numbers states that as the numbers of identically distributed and randomly generated variables increase, the sample mean tends to converge to the theoretical mean [125]. Law of large numbers is expressed as:

$$P(|\nu - \mu| > \epsilon) \to 0 \qquad as\ N \to \infty \tag{3.12}$$

Where ν is the theoretical mean, μ is the sample mean, and ϵ is the acceptable tolerance for deviation between ν and μ. Hence, in order to ensure consistency of ERM principle on convergence, it is necessary for empirical risk to converge to real risk **uniformly** over function $f(X)$. Uniform convergence for ERM is defined as:

$$P\left(\sup\left|R_{emp}(f) - R(f)\right| > \epsilon\right) \to 0 \qquad as\ N \to \infty \tag{3.13}$$

For large number of finite training samples (large N), convergence rate is fast if the following Hoeffding's Inequality holds true:

$$P(|S(g) - S_n(g)| > \epsilon) \leq 2e^{-2\epsilon^2 N} \tag{3.14}$$

where $g(X) = 0$ when $f(X) \neq Y$ and $g(X) = 1$ when $f(X) = Y$, $S(g)$ denotes the expected value of error for function g and $S_n(g)$ represents the frequency of error in n training set.

The bound of this rate of convergence is derived from here and is known as Vapnik Chervonenkis dimension (VC dimension) which will be discussed in the next section.

3.3.2 VC Dimension

Modern Learning Theory depends on bounding the difference between empirical and real risk of error uniformly over a set of hypothesis. Vapnik-Chervonenkis dimension (VC dimension) determines this bound and itself is a measure of complexity (capacity) of learning functions. VC dimension measures the largest number of examples that can be explained by the function. Vapnik [124] showed that with probability $1 - \eta$, the inequality

$$\underbrace{R(f)}_{Testing\ error} \leq \underbrace{R_{emp}(f)}_{Training\ error} + \underbrace{\sqrt{\frac{h\left(\ln(\frac{2N}{h})+1\right)-\ln\left(\frac{\eta}{4}\right)}{N}}}_{VC\ confidence} \quad (3.15)$$

where h is the VC dimension of the classifier and N is number of training examples and the inequality only holds true when (N>>h).

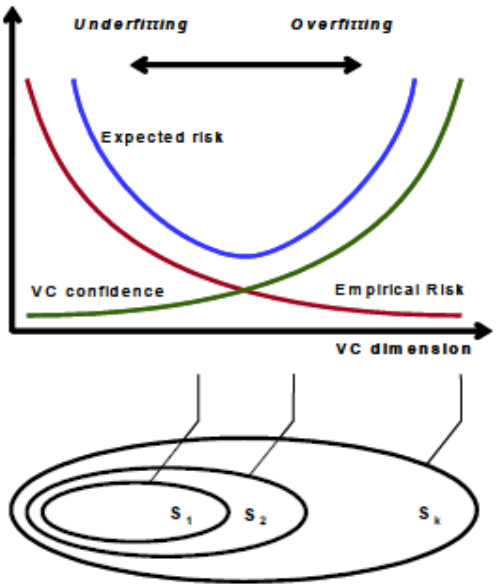


Figure 3.2 Concept of Structural Risk Minimization [126]

As can be seen from Figure 3.2, the higher the VC dimension, the bigger the capacity to learn but generalization capability goes down. Increase in VC dimension also leads to overfitting, this makes the learning algorithm become complex and requires a long time to execute. However, controlling the VC

dimension by reducing it leads to an increase in training error. Hence, a learning structure that decreases the VC dimension at the expense of the smallest possible increase in training error is needed. To solve the overfitting problem, complexity of the function class must be restricted. Occam's razor principle states that when there are competing hypotheses that make the same prediction, the simpler function that is able to explain most of the data is better [127]. This leads to concept of Structural Risk Minimization (SRM).

3.3.3 Structural Risk Minimization

Common problem faced in machine learning is overfitting, when a generalized model selected from a finite data set has poor performance on generalizing unseen data samples, because the model is tailored strongly to the training data. The concept of Structural Risk Minimization is to find an optimal model by striking a balance between empirical error and VC dimension. Principle of SRM is to construct nested structure for a group of function classes with increasing VC dimension. Then select the model that is minimal in VC confidence and empirical error. This can be visualized from Figure 3.2. From the figure, the model that is most likely to be chosen is S_2.

3.3.4 Support Vector Machines

Support Vector Machines are supervised learning classifiers that categorise new data by selecting the optimal hyperplane through the implementation of the concept of Structural Risk Minimization. By finding the optimal hyper-plane, the

SVM selects a learning structure which minimizes training error and generalization error at the same time. The VC dimension is related to the optimal hyper-plane by the upper bound imposed on it by the margin of separation between that hyper-plane and the closest data point. The structure with the smallest VC dimension corresponding to a still acceptable training error is chosen because the SVM picks the optimal hyper-plane. Vapnik showed that the VC dimension of a separating hyperplane with a margin m is bounded as follows:

$$h \leq min \left\lceil \frac{R^2}{m^2} \right\rceil, D + 1 \tag{3.16}$$

where D is the dimensionality of input space and R is the radius of the smallest sphere around the origin containing all input vectors. Hence, it is shown that minimizing VC dimension can be achieved by maximizing margin from the inequality. However, when given a set of linearly separable data $\{(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)\}, x \in R^D, y \in \{-1, +1\}$ which belongs to one of the two classes, hyperplane that is obtained need to be justified as to whether it is optimal [128].

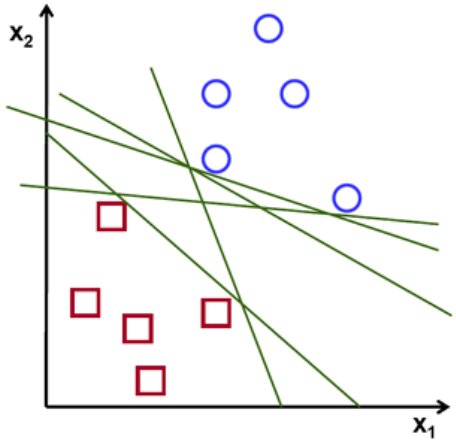


Figure 3.3 Which hyperplane is the optimal?[129]

The goal of SVM is to design classifier $g(x) = w^T x + b$ that maximizes the margin between two linearly separable classes [128]. The optimal hyperplane is obtained by maximizing margin of the training data.

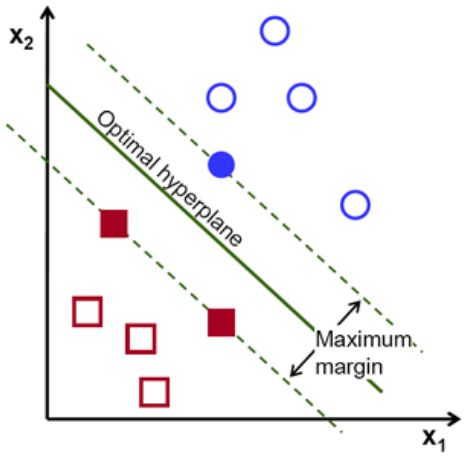


Figure 3.4 Optimal Hyperplane [129]

The definition of separating hyperplane is $g(x) = 0$. Each hyperplane is characterized by direction represented by weight vector, w^T and position in the space represented by the bias, b. For each direction, SVM chooses the hyperplane that leaves the same distance from the nearest points from each class. This made the margin twice the distance. For a binary linearly separable data, the function of hyperplane is 0. When the function is more than 0,

$$(w^T x_i) + b > 0 \tag{3.17}$$

it is classified as 1. When the function is less than 0, the data is classified as -1 class.

$$(w^T x_i) + b < 0 \tag{3.18}$$

This can be shown in diagram as following.

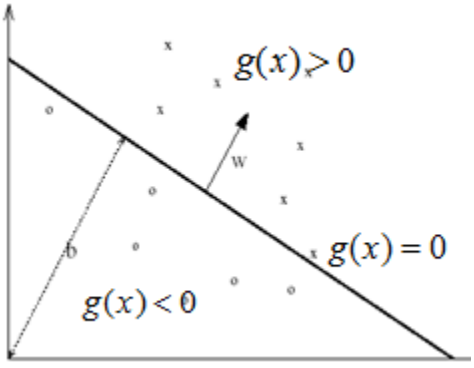


Figure 3.5 Function of Hyperplane[130]

The distance between point x from plane (w, b) is given by

$$\frac{|w^T x + b|}{\|w\|} \tag{3.19}$$

By scaling the weight and bias, the discriminant function at nearest points from each class is ± 1 .

$$|w^T x + b| = 1 \tag{3.20}$$

The distance from the closest data point to the boundary is

$$\frac{|w^T x + b|}{\|w\|} = \frac{1}{\|w\|} \tag{3.21}$$

Thus, the margin is given by

$$m = \frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|} \tag{3.22}$$

Minimizing $\|w\|$ to maximize the margin is equivalent to finding:

$$\min \frac{1}{2} \|w\|^2 \tag{3.23}$$

subject to $y_i(x_i w + b) - 1 \geq 0 \; \forall_i$. This term is used to perform Quadratic Programming optimization.

3.3.5 Quadratic Programming

A computationally efficient procedure to find optimal hyperplane for using training data is to be presented. The constrained minimization problem is solved by introducing Langrangian. The Primal Langrange Function is as following:

$$L_P \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L} \alpha_i [y_i(x_i \cdot w + b) - 1]$$

$$\equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L} \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^{L} \alpha_i \quad (3.24)$$

The solution to this optimization problem is determined by the saddle point of Langrangian Function which has to minimize w and b and maximize α by differentiating L_P with respect to w and b, then setting the derivative to zero. Hence, conditions of optimality are obtained:

$$\frac{\partial L_P}{\partial w} = 0 => w = \sum_{i=1}^{L} \alpha_i y_i x_i \quad (3.25)$$

$$\frac{\partial L_P}{\partial b} = 0 => \sum_{i=1}^{L} \alpha_i y_i = 0 \quad (3.26)$$

Dual Lagrange function derived from primal Langrange function:

$$L_D \equiv \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \quad (3.27)$$

Subject to $\alpha_i \geq 0 \; \forall_i, \; \sum_{i=1}^{L} \alpha_i y_i = 0$.

$$\equiv \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i H_{ij} \alpha_j \; where \; H_{ij} = \; y_i y_j x_i \cdot x_j$$

$$\equiv \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T H \alpha \tag{3.28}$$

Subject to $\alpha_i \geq 0 \; \forall_i, \sum_{i=1}^{L} \alpha_i y_i = 0$.

The dual form requires only dot products of each training data. This is important to Kernel Trick to be discussed later in Chapter 4.

The way of solving the problem is transforming from minimization of L_P to easier one by maximization of L_D (Note that L_D depends only on Lagrange multipliers).

$$\max_{\alpha} \left[\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T H \alpha \right] \tag{3.29}$$

Subject to $\alpha_i \geq 0 \; \forall_i, \sum_{i=1}^{L} \alpha_i y_i = 0$.

Maximization of the function can be achieved when α_i is positive. By referring to Equation (3.27), the following three conditions are considered:

Condition 1: If features x_i and x_j are completely different, dot products between these two features is 0. Hence,

$$L_D \equiv \sum_{i=1}^{L} \alpha_i$$

where only the first term on right hand side of Equation (3.27) contributes to L_D.

Condition 2: If features x_i and x_j are similar, dot products between these two features is 1 which leads to two sub conditions.

- Sub condition 1: when both x_i and x_j predicts the same output where $y_i = y_j = 1$ or -1, then $y_i y_j x_i \cdot x_j = 1$ and $\alpha_i \alpha_j y_i y_j x_i \cdot x_j > 0$, the value of L_D will be decreased.
- Sub condition 2: when both x_i and x_j predicts the opposite output where $y_i = -1$ and $y_j = 1$, then $y_i y_j x_i \cdot x_j = -1$ and $\alpha_i \alpha_j y_i y_j x_i \cdot x_j < 0$, L_D will be maximized. In this case, the feature vectors that predict the opposite output maximize the margin of separation.

3.3.6 Section Summary

Support Vector Machine controls the cost of misclassification parameter, C which avoids overfitting and allows for slight inaccuracy in training data. Kernel trick for SVM enable the building of kernel which suitable for the domain with low computational cost as it depends only on the support vectors. This has made SVM ideal for this research because of the inaccuracy induced by approximate models. Domain of this research focuses on identifying village with economic

potential in cultivating underutilised crops through generating data from approximate farm income models. Hence, kernel that derived from function describing income distribution within a group of people is engineered which suitable for domain of this research. This is covered in details in Section 4.6.

Chapter 4 Methodology

4.1 Introduction

This project focusses on trying to predict the potential of successfully implementing underutilised crops in rural village. This is accomplished by estimating the impact of growing underutilised crops on the income of the village. However, the lack of data from underutilised crops causes difficulty to train Support Vector Machine (SVM) for classification. Hence normal crop models are used to generate artificial data using Genetic Algorithm to train SVM. An experiment on the capability of SVM to classify, based on data generated from an 'approximate' model will be further discussed in Chapter 5 of this thesis. An approximate model here refers to a normal crop model that includes characteristics or variables that are approximately similar for an underutilised crop model. In this project, village characteristics (such as demographic of village, distance from village to nearest market and so on) are taken into account in order to predict the success of the village when it comes to growing under-utilized crops.

The methodology was designed in order to overcome the following research issues (statements):

Table 4.1 Methodology steps to solve Problem Statements

	Research Issues	What other people did	What this research proposes
1.	Lack of training data to train Support Vector Machine.	• As mentioned in Chapter 2, Li et al. [68] developed Genetic Algorithm-based Virtual Sample Generation to resolve the instability caused by using small data sets on back propagation Neural Network. • Jing et al. [28] generated virtual sample based on Gaussian distribution. This is done by generating random numbers following Gaussian distribution around the original sample data. More extensive of related studies have been discussed in previous chapter.	With the problem caused by insufficiency of real data for underutilised crops, normal crop model is used with Genetic Algorithm to generate artificial training data that fulfilled realistic boundaries.
2.	Economic models for underutilised crops are not available.	• Kantanantha et al. [131] developed a forecasting model on crop yield and price based on weather and economic growth. • Eravwoke et al. [97] studied major cash crop (oil palm, cotton, rubber, cocoa and groundnut) production in Nigeria and transport in relation to the economy of	Normal crop models that take into account village characteristics are used to generate sufficient training data to train Support Vector Machine to predict potential of successfulness of underutilised crop in a village.

		their country. More related studies have been covered in Chapter 2.	
3.	Inaccuracy of models	• Laura [109] demonstrated the superiority of Artificial Neural Networks over classical discriminative analysis by building a classification model to target market strategies to the identified group of consumers. • Mora-Lopez et al. [114] presented a hybrid model between multivariate regression model, artificial neural network and probabilistic finite automata on forecasting atmospheric clarity value.	Combine economic models and machine learning to complement each other in order to overcome limitations of a model and solve classification problems.
4.	Proper labelling of artificial training data generated from Genetic Algorithm.	• IBM is hiring people to only focus on data labelling manually due to the importance of labelled data as emphasized by the head of team who are creating Watson in IBM. Data labelling is important for machines to be able to learn even for unsupervised data as stated by Professor of AI at the University of New South Wales [70].	Automate data labelling based on metastable state of fitness value generated from GA.

The models that are described below were developed for normal crops, and may not be fully applicable to underutilised crops. As such, in order to reduce the error caused by applying the normal cash crop model to underutilised crops, a machine learning strategy was incorporated within the system. It is hypothesized that the SVM trained with model data is able to "filter" some of the noise associated with the normal crop data, thereby increasing the classification accuracy when applied to predicting the chances of success for rural villages in commercialising underutilised crops. A software experiment was carried out to prove this hypothesis.

4.2 Economic models

Most of the studies on economic models focused on the relationship between GDP growth and the independent variables that have been identified as significant. Minh studied the effects of demographic transition such as population growth on GDP of developing countries [132]. In a paper by Kantar et al. they discussed on relationships between electricity consumption and GDP [133]. Beyzatlar et al. explored relationship between transportation and GDP per capita [134]. Other studies used historical GDP to form GDP model in order to forecast future GDP at national level [135][136][137].

In this research, the approximate relationship of farm income to economic indicators for rural areas was used as the objective function for genetic algorithm. The relationship was found from different references [9][10][11].

As mentioned in Chapter 2, farm household income is listed as one of the commercialisation effects. Hence, successfulness of crops commercialisation can be measured through farm household income. Braun et al. [20] presented a conceptual framework on agricultural commercialisation. Main determinants of commercialisation are identified as population, demographic change, new technologies, new crops, infrastructure and market creation, macro economy and trade policy. The farm income models that are used to measure the commercialisation of crops to be presented in next section are chosen because the models fulfilled the main determinants of commercialisation.

4.2.1 Farm Income Models

Low farm income is one of the major consequences caused by low commercialisation from agricultural produce. This also contributes to poverty in the farm villages [20]. Farm income is a major contribution to economic development of a rural area in developing countries [138]. Hence, in order to gauge effect of commercialising underutilised crops on economic growth of rural area, models that relate farm income from crops to village characteristics are used in this project.

Multiple regression analysis was performed in these models to investigate the relationship between dependent variable (farm income) and multiple independent variables. General form of multiple regression equation can be expressed as follows:

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

Where y is the dependent variable, x_i $(i = 1,2 \ldots n)$ are the independent variables, a is the regression intercept and b_i $(i = 1,2 \ldots n)$ are the coefficients for each of the independent variables. This equation can also be used to predict the value of dependent variables based on the given values of independent variables. Multiple linear regression analysis involves several main assumptions for the prediction to be valid. It is assumed that the relationship between dependent and independent variables is linear. The variables are also assumed to be normally distributed. Homoscedasticity is assumed where the variance of error terms across the independent variables are constant. The three economic models selected are presented as following:

Model 1:

The model by Akimi [9] was used to investigate rice farm profitability in Kelantan, Malaysia by adopting Cobb Douglas production function.

log Y = log -0.3817 + 0.5629 log A + 0.1849 log B + 0.2778 log C + 0.0239 log D -0.1376 log E

Table 4.2 Details of Variables Involved in Model 1

	Variables	Unit	Description
Y	Gross income from rice	$	Income from rice production
A	Cultivated rice land value	$	Value of cultivated rice land
B	Total family labour inputs	$	Total wage rate of labor
C	Total material inputs such as seeds, pesticide and fertilizer	$	Total cost of material inputs
D	Farmer's ability index (Education, Experience, Knowledge)	Binary	
E	Damaged rice area (% to total planted acreage)	% to total planted acreage	Percentage of damaged rice acreage out of total planted acreage

Model 2:

Assis et al. [10] developed the model to study relationship on how socioeconomic factors affect income and productivity from pineapple farm in Samarahan, Sarawak.

Y = 0.049 *A + 0.546*B-0.253*C + 0.206*D +0.034*E -0.038*F -0.012*G +0.066*H + 1.141*I – 0.442*J +0.492*K

Table 4.3 Details of Variables Involved in Model 2

	Variables	**Unit**	**Description**
Y	Monthly income from Pineapple	RM/Month	Income from Pineapple farming
A	Age	Number	Age distribution in the village
B	Gender	Male/Female	Gender distribution in the village
C	Marital Status	Married/Others	Marital Status of villagers
D	Formal Education	Year	Formal education conducted in school
E	Experience	Year	Farming Experience in Pineapple
F	Extension (Meeting with Extension Agent)	Number	Number of meetings with extension agent last year

G	Distance	Kilometres	Distance between house and farm
H	Household Size	Number	Number of family members in each household
I	Secondary Occupation	Yes/No	Secondary Occupation other than farming
J	Land Status	Self-Owned/Other	Status of the farming land
K	Land Size	Acre	Size of the farming land

Model 3:

Hassan [11] investigated the farm income in Sudan with the following model.

Y = 4.5861*A + 3.2999*B + 0.081*C + 1.295*D -5.238*E +17.11*F-3.653*G

Table 4.4 Details of Variables Involved in Model 3

	Variables	**Unit**	**Description**
Y	Farm Income	Sudanese Pound	Income from farming
A	Household Age	Year	Age distribution in the village

B	Household Size	Number	Number of family members in each household
C	Labour	man days	Labour in the farm
D	Value of assets	Sudanese Pound	Income during dry season
E	Market Distance	Kilometres	Distance of nearest market from the farm
F	Operational Area	Mukhammas	Size of farm land
G	Weather Condition	Binary	Weather: Sunny, Cloudy, Rainy

4.2.2 Analysis on Economic Models

Table 4.5 Overview of Model 1,2 and 3

	1	2	3
Model	log Y = log -0.3817 + 0.5629 log A + 0.1849 log B + 0.2778 log C + 0.0239 log D -0.1376 log E	Y = 0.049 *A + 0.546*B- 0.253*C + 0.206*D +0.034*E - 0.038*F -0.012*G +0.066*H + 1.141*I – 0.442*J +0.492*K	Y = 4.5861*A + 3.2999*B + 0.081*C + 1.295*D -5.238*E +17.11*F-3.653*G
Type of Crops	Grain Crop	Fruit	General Food Crop
Countries of Analysis	Kelantan, Malaysia	Sarawak, Malaysia	Sudan
Variables	Y - Gross income from rice A - Cultivated rice land value B - Total family labour inputs	Y - Monthly income from Pineapple A - Age B - Gender	Y - Farm Income A - Household Age B - Household Size C - Labour

	C - Total material inputs such as seeds, pesticide and fertilizer D - Farmer's ability index (Education, Experience, Knowledge) E - Damaged rice acreage (% to total planted acreage)	C - Marital Status D - Formal Education E - Experience F - Extension (Meeting with Extension Agent) G - Distance H - Household Size I - Secondary Occupation J - Land Status K - Land Size	D - Value of assets E - Market Distance F - Operational Area G - Weather Condition

4.2.3 Analysis on Farm Income Models

Land value takes up a big part of assets value in agricultural sector. Duncan [139] states that value of land is determined by value of products produced on the land. Hence, increased in land value indicates that products value increases which positively affects farm income.

In general, crops yield increases as total agricultural inputs such as seeds, pesticide, fertilizer and labours increase [140]. In most regions that focus on agricultural sector, subsidies on agricultural inputs are provided to reduce the burden in terms of production cost for farmers [141].

As contrary to public beliefs, studies show that older farmers have positive impact to crops productivity which indicates positive effects on farm income. With experience and knowledge, older farmers are able to work more efficiently on their farms [142] [143]. .

It is found that farm income generated by women in an agricultural role is lower as compared to male counterparts. Study by Shambel [144] shows that most extension agents are men and only works with men which reduce the chances for women to access technology inputs that help in increasing farm productivity. Females in farms also allocate their time to household work while males only focus on farm work.

According to studies [145] [146], married farmers perform better in terms of farm productivity and farm income than unmarried farmers. Married farmers are more likely to produce more to generate more income to support the family. This leads to farmers accessing agricultural information to improve yields. Family labours available also aids in increasing output and income.

Improvement in education levels for farmers helps in increasing farm income as higher education enhances the chances of modernizing agriculture sector [147][148]. Hence, the education level of farmers positively impacts the farm income.

It is found that farmers with more farming experience are more likely to bring in higher income to the farm up to only a certain level [149][150].

Extension meeting is held aiming to bridge gap between agricultural research and experiments carried in laboratories to simulate actual farm conditions in order to increase farm productivity and income [151]. However, some studies shows negative impact of extension meetings in farm income like in Model 2, due to lack of guidance on new information or technology which causes productivity and income to decrease [152].

Distance between farm and house of farmers negatively impact the income because of the lack of proper road infrastructure, and transportation cost is added to higher production cost. [153]

Increase in household size leads to increase in farm income because household members are expected to supply the labour in farm production [154] [155].

Income from secondary occupations of farmers such as fishing, trading, and hunting helps to increase farm income [155].

By owning the land, farmers are able to make general and proper investment decisions on the land they own and this also leads to higher farm income results [156]. It also provides accessibility to credit opportunities which can further on improve land usage and results in even more income [157].

Land size is positively affecting the farm income because bigger land indicates higher yield. Larger land also makes it viable for farmers to use farm technology to cut down labour cost which make up the bigger part of production cost [158] [159].

When the market is easily accessible by farmers, they can sell their farm produce directly in the market. It reduces the transportation costs and eliminates the exploitation of middle men [160][161]. This also translates into higher farm income.

Weather conditions are critical factors for crop yield. Unfavourable weather conditions have negative impact on crop production and affect farm revenue [17] and cost of inputs, as bad weather is likely to wash away input materials [162].

4.2.4 Inputs variables to the models

The farm household income model is based on per capita income. Data collected from villages is required to pre-process and classify according to the types of variables. Level of Measurement describes the variable by classifying the variable into nominal, ordinal or interval-ratio [163] [164]. Central tendency will be used to summarize and represent the entire dataset for the variable. Commonly used central tendencies are mean, median and mode. Each of the central tendency measure is used based on the suitability of data for each level of measurement.

Table 4.6 Central Tendency Measures

Level of Measurement	Central Tendency	Variables
Nominal	Mode	Ability index, Gender, Marital Status, Secondary Occupation, Land Status, Weather Condition
Ordinal	Median	-
Interval-ratio	Mean for Symmetrical Distribution, Median for Skewed Distribution	Land value, Labour, Material inputs, Damaged crops land, Age, Education, Experience, Extension Meeting, Distance, Household size, Land

		size, Value of assets, Market Distance

4.2.5 Constraints

Constraints show the relationship between design variables and other design parameters. Inequality constraints define a design variable as smaller than, larger than or equal to a resource value. [20] GA is used in this project because it is possible to set constraints and boundaries that reflect the real life situation. Upper boundaries and lower boundaries for each of the variables are justified through literature.

1. Household Size
    - In a study on China's ethnic minority populations, average household size of Yunnan villages is estimated at 4 persons per household. [165]
    - Mean household size of rural Pakistan, rural Punjab and rural Balochistan is 5 to 8 members per household whilst rural Sindh is between 4 to 9 persons each household. [152]
    - Rural Cambodia has average household size of around 5 people which is shown in a study on rural development of Cambodia. [166]

- Pasanen did an analysis on household and village levels in Laos where the distribution of household size in Laos is between 1 to 22 people per household. [167]

Hence, it can be concluded that the lower bound and upper bound for household size in rural village is 1 and 22 respectively.

2. Age
    - Life expectancy is at 71.4 year old. [168]
3. Age of Household Head
    - Age of Household Head is expected to have the same life expectancy as other villagers (as above) [168].
4. Formal Education
    - India's official schooling entry age is at 5 (primary school) till age 19 (secondary school).
    - Malaysia adapts 6-3-2-2 formal education system, starting at the age of 6 to age of 17. [169]
    - In Nigeria, formal education includes 6 years primary education, 3 years junior secondary education, 3 years senior secondary education and 4 years higher education [170]. However, it is also shown that 40% of children of 6 to 11 years old in Nigeria did not attend school [171].

It is safe to conclude that lower bound for formal education is 0 considering some of the children did not accept any formal education. Upper bound is 16 years.

5. Land Size

- In India, average farm size is between 1.6 ha (in Orissa) to 3.1 ha (in Chattisgarh). [172]
- Mean farm size of villages in Philippines ranges from 0.8 to 4.3 ha [173].
- Average farm size of Malawi ranges between 0.58 and 1.76 ha for some of the villages [174].
- Minimum and maximum farm size according to Food and Agriculture Organization (FAO) in Kenya, Ethiopia, Malawi, Niger, Nigeria, Tanzania, Uganda, Bangladesh, Nepal, Vietnam, Bolivia, Guatemala, Nicaragua and Albania are presented in the following table [175].

Table 4.7 Minimum and Maximum Farm Size in Kenya, Ethiopia, Malawi, Niger, Nigeria, Tanzania, Uganda, Bangladesh, Nepal, Vietnam, Bolivia, Guatemala, Nicaragua and Albania

Area	Minimum Farm Size (ha)	Maximum Farm Size (ha)
Kenya	0.04	8.9
Ethiopia	0	18
Malawi	0.004	9.51

Niger	0.0002	34
Nigeria	0.00002	80.22
Tanzania	0.001	21
Uganda	0.004	8.5
Bangladesh	0.004	2
Nepal	0.0034	16.5
Vietnam	0.001	12
Bolivia	0.0003	151
Guatemala	0	14.33
Nicaragua	0.2	282
Albania	0.003	3.34

The lower bound and upper bound for land size should be 0 and 80.22 respectively.

6. Market Distance

    - In Vietnam, the distance between nearest market and villages is presented in a study of on-farm conservation as listed below [176]:

Table 4.11 Market Distance of Villages in Vietnam

Villages	**Distance to nearest market (km)**
Cang village	18
Tat village	22
Qungmao village	5
Yen minh village	5
Dong lac village	2
Kien thanh village	1.5
Dai An village	40

- Meanwhile in Philippines, the distance of primary market to villages is as follows [177]:

Table 4.8 Market Distance of Villages in Philippines

Villages	Distance to nearest market (km)
Isabela	7-10
Camarines Sur	2-15
Cebu	4-8
Leyte	1.5-10
Bukidnon	5-57
South Cotabato	4-8
Cotabato	18-21

- Distance of some of the villages in India from the nearest output market is tabulated in the table below [178]:

Table 4.9 Market Distance of Villages in India

Villages	Distance to nearest output markets
Amudur	10
Duli	5
Kalpattu	8.6
Meppathurai	6.4
Nesal	6.5
Sirungathur	7
Vayalur	6.4
Veerasambanur	7
Vegamangalam	6
Vengodu	10
Vinayagapuram	13

Lower bound and upper bound of nearest market distance from the village is at 1.5 and 40 respectively.

7. Farm Distance
    - Agricultural guideline for Botswana states that distance between farm and village is recommended to be 8-19km for big villages to grow and 5km between farm and village for small villages [179].
8. Experience
    - It is assumed that maximum experience of farming starting at the age of 6 till age 71.4 is 65.4 years.
9. Total Material Inputs (seeds, pesticide and fertilizer)
    - Material inputs costs per acre of rice production in US are as follows [180]:

Table 4.10 Material Input Costs Per Acre of Rice Production in US

Materials	Costs per Acre
Seed	83.44
Fertilizer	130.25
Chemicals	115.65
Total	354.09

    - In Indonesia, calculation of total material inputs used per acre is presented in the table below [181]:

Table 4.11 Material Input Costs Per Hectare in Indonesia

Materials	**kg per hectare**	**Rupiahs per kg**	**Rupiahs per hectare**
Fertilizer	400	125	50,000
Liquid Insecticide	2 (liters per hectare)	3000 (Rupiahs per kg)	6000
Granulated Insecticide	8	700	5600
Seeds	25	400	10,000

10. Labours

- It is stated that labour cost is from $26.95 to $31.35 per acre in Iowa, United States depending on the type of crop that is planted [182].
- Daily wages in Indonesia for rice production ranges between Rp1200 ($18) per day for male and Rp700 ($10.50) per day for female [181].

11. Crops Land Value

- Land value in Iowa is approximately $266 per acre [182].

12. Number of meetings with extension agent per year

- Assume that if meeting is held at maximum once per week, it will be held for 54 times per year whilst if the meeting only held at once per month, it will be held for 12 times per year.

13. Value of Assets

- It is reported that the average value of farm assets owned by rural household is approximately $15,000 in India [183].
- In Thailand, village household assets for small non mechanized village household asset is estimated at $490 while large mechanized village assets at $2900 [184].

4.2.6 Summary of Variables

Table 4.12 Summary of Lower and Upper Bounds for Each of the Variables

Variables	Lower Bound, L_x	Upper Bound, U_x
Cultivated crops land value	0	266
Total family labour inputs	10.50	31.35
Total material inputs such as seeds, pesticide and fertilizer	13	700
Damaged crops acreage (percentage)	0	100
Age	1	71.4
Gender	1	2
Marital Status	1	2
Formal Education	0	16
Experience	0	65.4
Number of meeting with extension agent per year	12	54
Distance (between house and farm)	5	19
Household Size	1	22
Secondary Occupation	1	2
Land Status	1	2
Land Size	1	81

Labour	10.50	31.35
Value of Assets	1	15,000
Market Distance	1.5	40
Weather Condition	1	3
Age of Head of Household	1	71.4

4.3 Artificial Training Data Generated by Genetic Algorithm

The proposed process to generate artificial training data is as presented in Figure 4.1 and is described in the following:

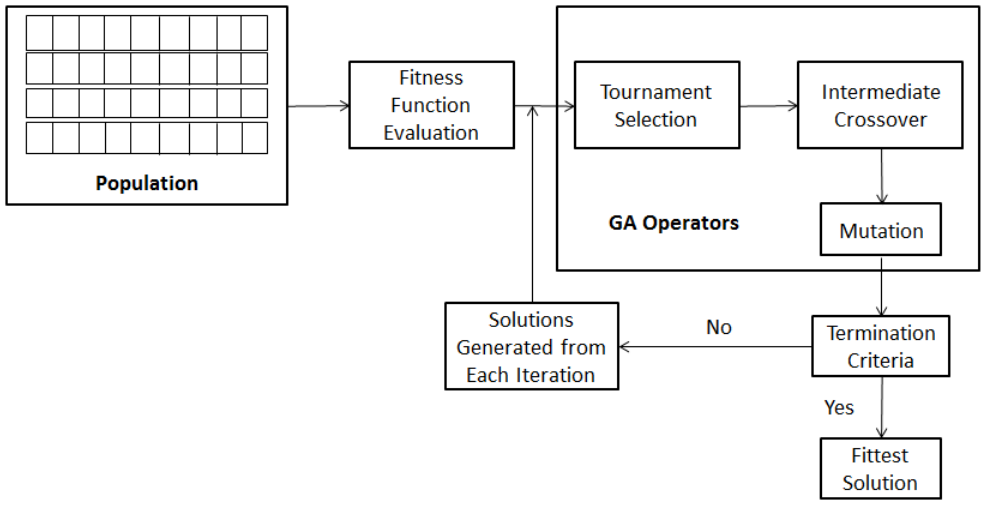


Figure 4.1 Genetic Algorithm Artificial Data Generation Process

1. Population initialization: Random population that follows uniform distribution is initialized.

2. Fitness function evaluation: Equation of Farm income model that relates crops income to village characteristics as mentioned in Section 4.3 is adapted as fitness function, $f(x)$. Fitness of each individual s_n (where $n = 1,2,3, \dots, M$) in the population is evaluated. Fitness score, $F(s_n)$ is generated based on the model and constraints that are defined.

3. Genetic Algorithm Operators: Genetic Algorithm operators consist of three processes: selection, crossover and mutation.

i. Tournament selection selects individuals randomly from the population. Individual with highest fitness value (fittest) from each round of tournament is kept. Eventually, fitness of the population is improved after each successive generation.

ii. Intermediate crossover takes place on a pair of selected chromosomes P_1 and P_2. Intermediate crossover is governed by Ratio as follows:

$$C = P_1 + \alpha \cdot Ratio \cdot (P_2 - P_1) \quad (4.1)$$

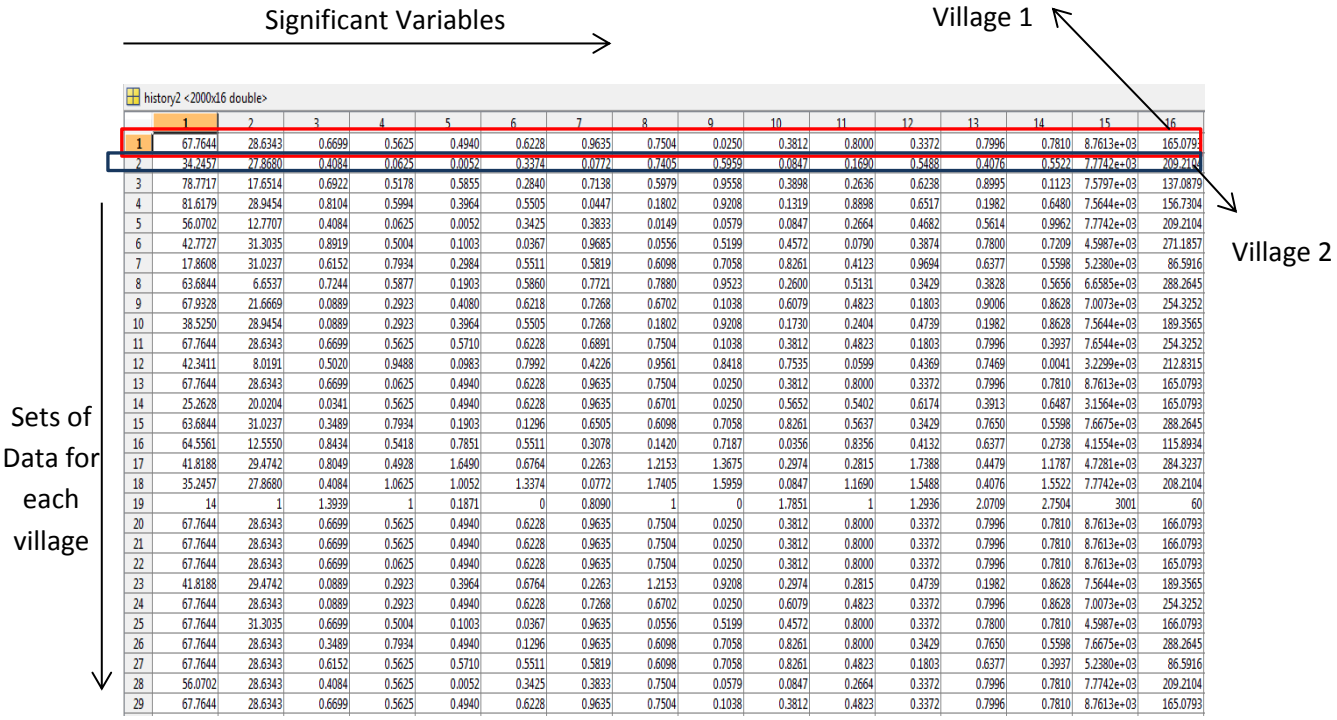
where C is the child created through crossover and α is random number between 0 to 1. The value of Ratio can be scalar or vector of the length of number of variables.

iii. Process of mutation randomly changes a gene of chromosomes in the population to create a mutated child. This is to avoid current solution from falling into local maximum solution.

4. Termination Criteria: After each iteration of fitness values computation, the following criterion is checked to decide whether the fitness value improved. If termination criteria are fulfilled, the process is terminated.

$$F(s_n) - F(s_{n-1}) \leq 0.0000001 \quad (4.2)$$

4. Solutions from previous iterations: Every set of solutions from successive generation is stored. After the process is terminated, these solutions are used as generated training data.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	67.7644	28.6343	0.6699	0.5625	0.4940	0.6228	0.9635	0.7504	0.0250	0.3812	0.8000	0.3372	0.7996	0.7810	8.7613e+03	165.0793
2	34.2457	27.8680	0.4084	0.0625	0.0052	0.3374	0.0772	0.7405	0.5959	0.0847	0.1690	0.5488	0.4076	0.5522	7.7742e+03	209.2104
3	78.7717	17.6514	0.6922	0.5178	0.5855	0.2840	0.7138	0.5979	0.9558	0.3898	0.2636	0.6238	0.8995	0.1123	7.5797e+03	137.0879
4	81.6179	28.9454	0.8104	0.5994	0.3964	0.5505	0.0447	0.1802	0.9208	0.1319	0.8898	0.6517	0.1982	0.6480	7.5644e+03	156.7304
5	56.0702	12.7707	0.4084	0.0625	0.0052	0.3425	0.3833	0.0149	0.0579	0.0847	0.2664	0.4682	0.5614	0.9962	7.7742e+03	209.2104
6	42.7727	31.3035	0.8919	0.5004	0.1003	0.0367	0.9685	0.0556	0.5199	0.4572	0.0790	0.3874	0.7800	0.7209	4.5987e+03	271.1857
7	17.8608	31.0237	0.6152	0.7934	0.2984	0.5511	0.5819	0.6098	0.7058	0.8261	0.4123	0.9694	0.6377	0.5598	5.2380e+03	86.5916
8	63.6844	6.6537	0.7244	0.5877	0.1903	0.5860	0.7721	0.7880	0.9523	0.2600	0.5131	0.3429	0.3828	0.5656	6.6585e+03	288.2645
9	67.9328	21.6669	0.0889	0.2923	0.4080	0.6218	0.7268	0.6702	0.1038	0.6079	0.4823	0.1803	0.9006	0.8628	7.0073e+03	254.3252
10	38.5250	28.9454	0.0889	0.2923	0.3964	0.5505	0.7268	0.1802	0.9208	0.1730	0.2404	0.4739	0.1982	0.8628	7.5644e+03	189.3565
11	67.7644	28.6343	0.6699	0.5625	0.5710	0.6228	0.6891	0.7504	0.1038	0.3812	0.4823	0.1803	0.7996	0.3937	7.6544e+03	254.3252
12	42.3411	8.0191	0.5020	0.9488	0.0983	0.7992	0.4226	0.9561	0.8418	0.7535	0.0599	0.4369	0.7469	0.0041	3.2299e+03	212.8315
13	67.7644	28.6343	0.6699	0.0625	0.4940	0.6228	0.9635	0.7504	0.0250	0.3812	0.8000	0.3372	0.7996	0.7810	8.7613e+03	165.0793
14	25.2628	20.0204	0.0341	0.5625	0.4940	0.6228	0.9635	0.6701	0.0250	0.5652	0.5402	0.6174	0.3913	0.6487	3.1564e+03	165.0793
15	63.6844	31.0237	0.3489	0.7934	0.1903	0.1296	0.6505	0.6098	0.7058	0.8261	0.5637	0.3429	0.7650	0.5598	7.6675e+03	288.2645
16	64.5561	12.5550	0.8434	0.5418	0.7851	0.5511	0.3078	0.1420	0.7187	0.0356	0.8356	0.4132	0.6377	0.2738	4.1554e+03	115.8934
17	41.8188	29.4742	0.8049	0.4928	1.6490	0.6764	0.2263	1.2153	1.3675	0.2974	0.2815	1.7388	0.4479	1.1787	4.7281e+03	284.3237
18	35.2457	27.8680	0.4084	1.0625	1.0052	1.3374	0.0772	1.7405	1.5959	0.0847	1.1690	1.5488	0.4076	1.5522	7.7742e+03	208.2104
19	14	1	1.3939	1	0.1871	0	0.8090	1	0	1.7851	1	1.2936	2.0709	2.7504	3001	60
20	67.7644	28.6343	0.6699	0.5625	0.4940	0.6228	0.9635	0.7504	0.0250	0.3812	0.8000	0.3372	0.7996	0.7810	8.7613e+03	166.0793
21	67.7644	28.6343	0.6699	0.5625	0.4940	0.6228	0.9635	0.7504	0.0250	0.3812	0.8000	0.3372	0.7996	0.7810	8.7613e+03	166.0793
22	67.7644	28.6343	0.6699	0.0625	0.4940	0.6228	0.9635	0.7504	0.0250	0.3812	0.8000	0.3372	0.7996	0.7810	8.7613e+03	165.0793
23	41.8188	29.4742	0.0889	0.2923	0.3964	0.6764	0.2263	1.2153	0.9208	0.2974	0.2815	0.4739	0.1982	0.8628	7.5644e+03	189.3565
24	67.7644	28.6343	0.0889	0.2923	0.4940	0.6228	0.7268	0.6702	0.0250	0.6079	0.4823	0.3372	0.7996	0.8628	7.0073e+03	254.3252
25	67.7644	31.3035	0.6699	0.5004	0.1003	0.0367	0.9635	0.0556	0.5199	0.4572	0.8000	0.3372	0.7800	0.7810	4.5987e+03	166.0793
26	67.7644	28.6343	0.3489	0.7934	0.4940	0.1296	0.9635	0.6098	0.7058	0.8261	0.8000	0.3429	0.7650	0.5598	7.6675e+03	288.2645
27	67.7644	28.6343	0.6152	0.5625	0.5710	0.5511	0.5819	0.6098	0.7058	0.8261	0.4823	0.1803	0.6377	0.3937	5.2380e+03	86.5916
28	56.0702	28.6343	0.4084	0.5625	0.0052	0.3425	0.3833	0.7504	0.0579	0.0847	0.2664	0.3372	0.7996	0.7810	7.7742e+03	209.2104
29	67.7644	28.6343	0.6699	0.5625	0.4940	0.6228	0.9635	0.7504	0.1038	0.3812	0.4823	0.3372	0.7996	0.7810	8.7613e+03	165.0793

Figure 4.2 Example of Data Generated from Genetic Algorithm

Despite that Genetic Algorithm being frequently used to generate test data for software testing [8] [9] [10], the idea of generating training data for SVM purpose using Genetic Algorithm is not common as mentioned in previous chapter. In this case, the goal of GA is to look for all the possible solutions that satisfy fitness function, instead of only a single optimal solution. The fitness function of GDP model is constructed for GA to search for all solutions that fit the model. Each iteration of chromosomes population is retrieved to form a set of training data. Figure 4.2 shows the data generated from Genetic Algorithm. Each column

represents the significant variables in the fitness function. Each row shows the relative value of each village characteristics. Note that these data of villages are not real villages, but they reflect real life by setting proper constraints.

Constraints show the relationship between design variables and other design parameters. Constraints are not in any particular format due to their variation, depending on the problem and user. Generally, there are two types of constraints, which are equality and inequality. Inequality constraints define a design variable smaller than, larger than or equal to a resource value. For example, $x < 5$ is an inequality constraint where x is the design variable and 5 is the resource value. The equality constraints define the relationship between design variable and resource value which must be equal to each other. $x = 5$ is an example of equality constraint. [20]

As mentioned, previously GA is used in this project because it is possible to set constraints and boundaries that reflect the real life situation. For example, the boundary for Years of Education should be between 0 to 19 years, as one could be not educated at all or educated for up to 19 years from kindergarten to Ph.D. Details of this has been presented in previous section of this chapter.

Synthetic Data Generation

1. Initialize population

2. Calculate the fitness of population

3. Select elite parents from population

4. Perform crossover on parents to create new population

5. Perform mutation on the population

6. Update new fitness of population

7. Repeat 3 to 6 until converges

8. Retrieve all the generations as artificial data

4.4 Data Labelling

As discussed in previous chapter, Schema Theorem is often applied to explain convergence of Genetic Algorithm to find optimal solution. In this research, this characteristic is used to label data generated from GA. The point where the fitness value converges to optimal solution, is called 'breakpoint'. The breakpoint is where the data is labelled as 1 if above breakpoint, otherwise 0. This is applicable because at that point, the solution is converging to solutions that fit the fitness function. These solutions are data that will be used to train SVM as mentioned previously. Less fit solutions before convergence is data that are labelled as 0. Fit solutions approaching convergence are data to be labelled as 1.

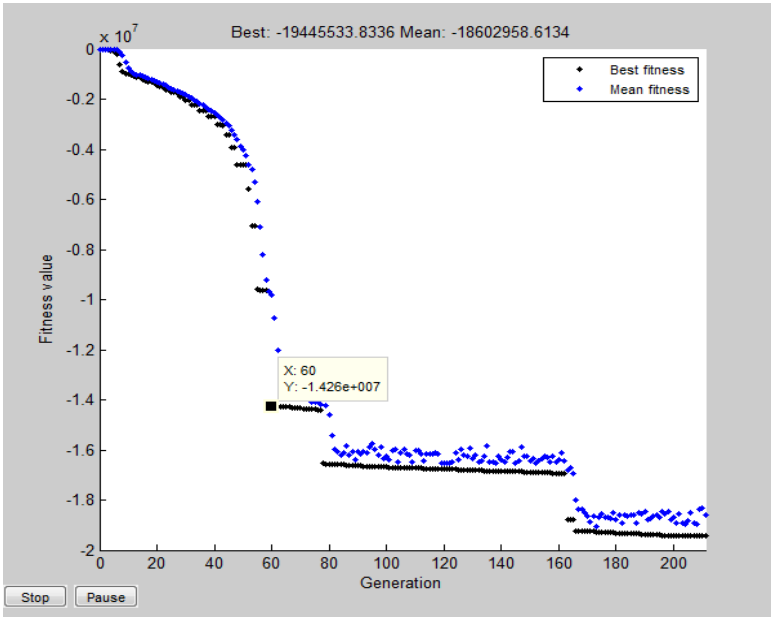


Figure 4.3 Breakpoint Label Example 1

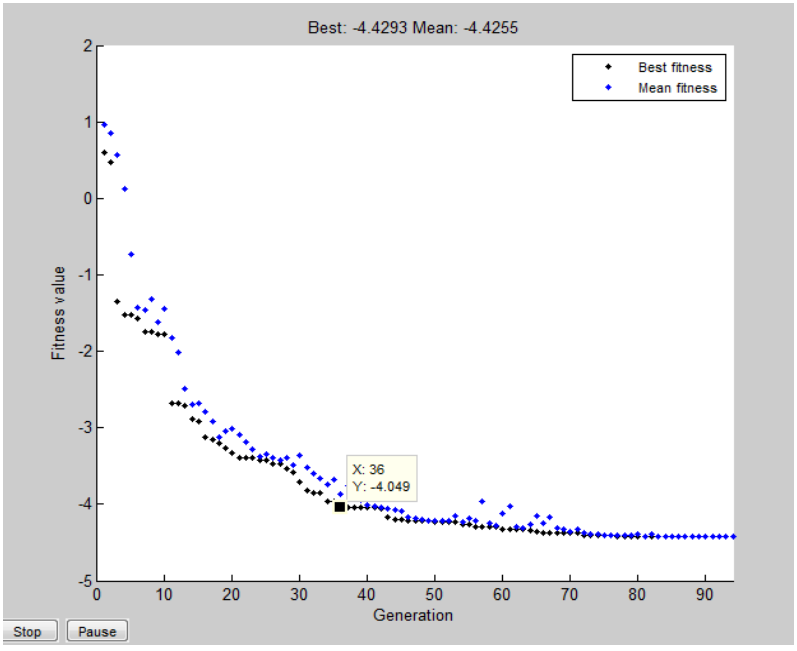


Figure 4.4 Breakpoint Label Example 2

4.5 Kernel Trick

As mentioned in Section 3.3.6, SVM kernels which are suitable for domain application can be designed through kernel trick with low computational cost as it depends only on the support vectors. So far, the classification problem mentioned involves only linearly separable data. There are many classification problems that are not linearly separable in space of inputs x as shown in Figure 4.5 (Left). In this case, applying a fixed non-linear mapping of the data to a higher dimensional feature space in Figure 4.5 (Right) with a suitable mapping function $x \rightarrow \varphi(x)$ is needed. In Figure 4.5, data is mapped from two dimensional space to three dimensional space:

$$\varphi: \mathbb{R}^2 \mapsto \mathbb{R}^3$$

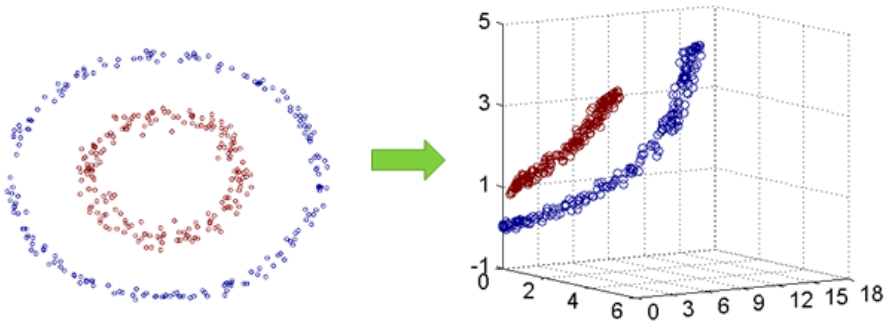


Figure 4.5 Mapping nonlinearly separable data to higher space [185]

The nonlinear machine of SVM can be built in two steps:

1. Mapping the nonlinear input vector into high dimensional feature space.
2. Constructing an optimal hyperplane of the Support Vector Machine to classify data in feature space.

The important property of SVM is that it is represented in dual form which depends solely on the dot products between testing and training points. This means that it is possible to directly compute the inner product of $\langle \boldsymbol{\varphi}(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}') \rangle$ in feature space as function of original input data. This function is known as ***kernel function***. This can be expressed as:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\varphi}(\boldsymbol{x}) \cdot \boldsymbol{\varphi}(\boldsymbol{x}') \rangle \tag{4.3}$$

Adopting kernel function in order to avoid carrying out φ explicitly is known as the ***kernel trick***. The basic requirements to define a kernel $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is

1. K is symmetric where $K(x, x') = K(x', x)$.
2. K is positive semi-definite.

4.5.1 Mercer's Theorem

Mercer's Theorem characterise a symmetric positive semi-definite kernel where the Gram matrix (Kernel Matrix) must be positive semi-definite [21].

$$\int K(x, x')f(x)f(x')\,dx\,dx' \geq 0 \tag{4.4}$$

For all functions $f(x), f(y)$ satisfying the inequality

$$\int f^2(x)\,dx \leq \infty. \tag{4.5}$$

Different learning machines with arbitrary types of nonlinear decision surfaces can be constructed by using different expressions for inner products for nonlinear vector in input space. The three commonly used of inner product kernels for Support Vector Machines are:

Table 4.13 Kernels

Kernel	Function
Linear Kernel	$K(x, x') = x^T x'$
Polynomial Kernel	$K(x, x') = (x^T x' + 1)^d$
Radial Basis Kernel	$K(x, x') = e^{-\gamma\|x-x'\|^2}$

4.5.2 Wavelet Kernel

Wavelet theory has been widely used in signal processing applications. A wavelet starts at zero amplitude, follows by oscillations and returns back to zero amplitude. Some common wavelets (Meyer, Mexican Hat, and Shannon) are shown in Table 4.14. In wavelet analysis, functions are expanded in the form of translations and dilations of mother wavelet. Mother wavelet is defined as:

$$\psi_{a,c}(x) = |a|^{\frac{1}{2}} \cdot \psi(\frac{x-c}{a}) \tag{4.6}$$

where a is dilation factor, c is translation factor and $\psi_{a,c}(x)$ is the mother wavelet. Wavelet transform of a function $f(x) \in L^2(\mathbb{R})$ can be written as

$$W_{a,c}(f) = \langle f(x), \psi_{a,c}(x) \rangle \tag{4.7}$$

That is the dot product in $L^2(\mathbb{R})$ space.

The dot product of wavelet kernel function is

$$K(x, x') = \prod_{i=1}^{n} \psi(\frac{x_i - c_i}{a_i}) \psi(\frac{x'_i - c_i}{a_i}) \tag{4.8}$$

Dot product wavelet kernels are valid kernels by proving the following:

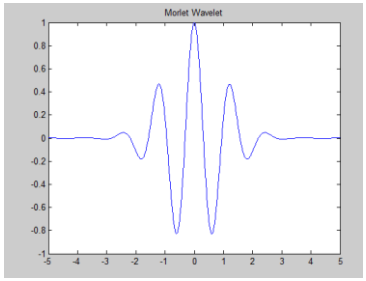
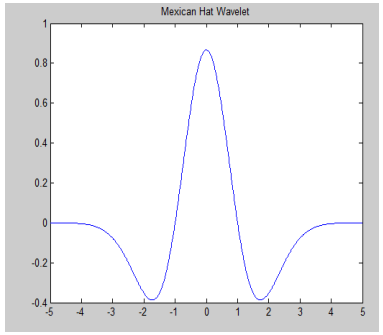
$$\begin{aligned} &\int K(x, x') f(x) f(x')\, dx\, dx' \\ &= \int \prod_{i=1}^{n} \psi\left(\frac{x_i - c_i}{a_i}\right) f(x) dx \int \prod_{i=1}^{n} \psi\left(\frac{x'_i - c_i}{a_i}\right) f(y) dy \\ &= \left(\int \prod_{i=1}^{n} \psi\left(\frac{x_i - c_i}{a_i}\right) f(x) dx\right)^2 \geq 0 \end{aligned} \tag{4.9}$$

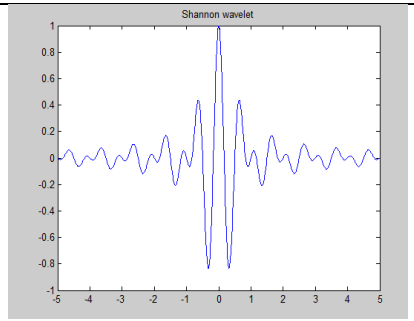
Hence, the valid translation-invariant wavelet kernel is

$$K(x, x') = \prod_{i=1}^{n} \psi(\frac{x_i - x'_i}{a}) \tag{4.10}$$

Various wavelet kernels has been proved and utilised in [186][187][188].

Table 4.14 Wavelet functions and corresponding mother wavelet and kernel functions

Wavelet	Function	Mother Wavelet	Wavelet Kernel Function
Morlet		$\psi(z - z') = \cos(wz) \cdot e^{-\frac{z^2}{2}}$	$K(x, x') = \prod_{i=1}^{n} \cos(w \frac{(x_i - x'_i)}{a_i}) \cdot e^{-\frac{\|x_i - x'_i\|^2}{2a_i^2}}$
Mexican Hat		$\psi(z - z') = (1 - z^2) \cdot e^{-\frac{z^2}{2}}$	$K(x, x') = \prod_{i=1}^{n} 1 - (\frac{(x_i - x'_i)^2}{a_i^2}) \cdot e^{-\frac{\|x_i - x'_i\|^2}{2a_i^2}}$

Shannon		$\psi(z - z') = 2\,\mathrm{sinc}\,2(z)$ $- \mathrm{sinc}(z)$	$K(x, x') = \prod_{i=1}^{n} 2\,\mathrm{sinc}\,2\left(\frac{x_i - x'_i}{a_i}\right)$ $- \mathrm{sinc}\left(\frac{x_i - x'_i}{a_i}\right)$

A translation invariant kernel $K(x, x') = K(x - x')$ is an admissible Support Vector kernel if and only if the Fourier transform of the kernel is non negative.

$$F[K](\omega) = (2\pi)^{-N/2} \int e^{-i\omega x}\ K(x)\ dx\ \geq 0 \tag{4.11}$$

Where $K(x) = 2\,\mathrm{sinc}(2x) - \mathrm{sinc}(x)$. Hence, the following is to be proven.

$$(2\pi)^{-N/2} \int e^{-i\omega x}\ [2\,\mathrm{sinc}(2x) - \mathrm{sinc}(x)]\ dx\ \geq 0 \tag{4.12}$$

Where $\mathrm{sinc}(x) = \frac{\sin x}{x}$.

By dividing $K(x)$ into two parts,

$$K(x) = \varphi_1(x) - \varphi_2(x) \tag{4.13}$$

Where $\varphi_1(x) = 2\,\mathrm{sinc}(2x)$ and $\varphi_2(x) = \mathrm{sinc}(x)$

Fourier transform of $K(x)$ begins with $\varphi_2(x)$:

$$\int_{-\infty}^{\infty} e^{-i\omega x}\,\mathrm{sinc}(x)\,dx = \int_{-\infty}^{\infty} e^{-i\omega x}\,\frac{\sin(x)}{x}\,dx$$

$$= \int_{-\infty}^{\infty} e^{-i\omega x}\,\frac{e^{ix} - e^{-ix}}{2ix}\,dx$$

$$= \frac{1}{2i}\left(\int_{-\infty}^{\infty} \frac{e^{i(1-\omega)x}}{x}\,dx - \int_{-\infty}^{\infty} \frac{e^{-i(1+\omega)x}}{x}\,dx\right) \tag{4.14}$$

Equation (4.14) is further subdivided into another two parts, where $\varphi'_1(x) = \int_{-\infty}^{\infty} \frac{e^{i(1-\omega)x}}{x} dx$ and $\varphi'_2(x) = \int_{-\infty}^{\infty} \frac{e^{-i(1+\omega)x}}{x} dx$. Let $z = (1-\omega)x$, $\frac{dz}{dx} = (1-\omega)$. Substitutes the expression into $\varphi'_1(x)$.

$$\varphi'_1(x) = \int_{-\infty}^{\infty} \frac{e^{iz}}{z} dz \tag{4.15}$$

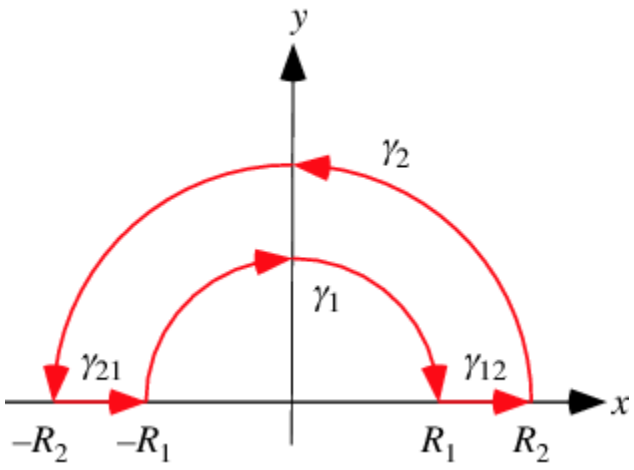


Figure 4.6 Path of contour integration [189]

By using contour integration and referring to Figure 4.6, $\varphi'_1(x)$ can be integrated as following:

$$\oint \frac{e^{iz}}{z} dz = \int_{\gamma_1} \frac{e^{iz}}{z} dz + \int_{\gamma_{12}} \frac{e^{iz}}{z} dz + \int_{\gamma_2} \frac{e^{iz}}{z} dz + \int_{\gamma_{21}} \frac{e^{iz}}{z} dz$$

$$= \int_{-R_1}^{R_1} \frac{e^{iz}}{z} dz + \int_{R_1}^{R_2} \frac{e^{iz}}{z} dz + \int_{R_2}^{-R_2} \frac{e^{iz}}{z} dz + \int_{-R_2}^{-R_1} \frac{e^{iz}}{z} dz$$

$$= \int_{\pi}^{0} \frac{e^{iz}}{z} dz + \int_{0}^{\infty} \frac{e^{iz}}{z} dz + \int_{0}^{\pi} \frac{e^{iz}}{z} dz + \int_{\infty}^{0} \frac{e^{iz}}{z} dz \tag{4.16}$$

Let $z = Re^{i\theta}$, $dz = iRe^{i\theta} d\theta$ on arc.

$$= \int_{\pi}^{0} \frac{e^{iR_1 e^{i\theta}}}{R_1 e^{i\theta}} iR_1 e^{i\theta} d\theta + \int_{0}^{\infty} \frac{e^{iz}}{z} dz + \int_{0}^{\pi} \frac{e^{iR_2 e^{i\theta}}}{R_2 e^{i\theta}} iR_2 e^{i\theta} d\theta + \int_{\infty}^{0} \frac{e^{iz}}{z} dz \tag{4.17}$$

By applying Jordan's Lemma, the third term is 0 as $R_2 \to \infty$.

$$= -\int_{0}^{\pi} id\theta + \int_{0}^{\infty} \frac{e^{iz}}{z} dz + 0 + \int_{\infty}^{0} \frac{e^{iz}}{z} dz$$

$$= -i\pi + \int_{-\infty}^{\infty} \frac{e^{iz}}{z} dz \tag{4.18}$$

By rearranging the above equation:

$$\int_{-\infty}^{\infty} \frac{e^{iz}}{z} dz = i\pi \tag{4.19}$$

When $1 - \omega > 0$, also defined as $\omega < 1$,

$$\frac{1}{2i} \int_{-\infty}^{\infty} \frac{e^{i(1-\omega)x}}{x} dx = \frac{\pi}{2} \tag{4.20}$$

For $\varphi'_2(x)$, flip the contour about axis and do a similar analysis as above, we obtain

$$= \int_{R_1}^{-R_1} \frac{e^{iz}}{z} dz + \int_{-R_1}^{-R_2} \frac{e^{iz}}{z} dz + \int_{-R_2}^{R_2} \frac{e^{iz}}{z} dz + \int_{R_2}^{R_1} \frac{e^{iz}}{z} dz \tag{4.21}$$

$$\frac{1}{2i} \int_{-\infty}^{\infty} \frac{e^{i(1+\omega)x}}{x} dx = \frac{\pi}{2} \tag{4.22}$$

for $\omega > 1$. Similarly, when $1 + \omega > 0$, which is also defined as $\omega > -1$

$$\frac{1}{2i} \int_{-\infty}^{\infty} \frac{e^{-i(1+\omega)x}}{x} dx = \frac{\pi}{2} \tag{4.23}$$

And $\omega < -1$

$$\frac{1}{2i} \int_{-\infty}^{\infty} \frac{e^{-i(1+\omega)x}}{x} dx = -\frac{\pi}{2} \tag{4.24}$$

Finally, the Fourier transform of $\varphi_1(x)$ is expressed as:

$$F(\omega) = \int_{-\infty}^{\infty} e^{-i\omega x} \operatorname{sinc}(x)\, dx \begin{cases} 0 & \omega < -1 \\ \pi & -1 < \omega < 1 \\ 0 & \omega > 1 \end{cases} \tag{4.25}$$

For $\varphi_2(x)$, using the same analysis,

$$\varphi_2(x) = \frac{1}{i} \left(\int_{-\infty}^{\infty} \frac{e^{i2(1-\omega)x}}{x} dx - \int_{-\infty}^{\infty} \frac{e^{-i2(1+\omega)x}}{x} dx \right) \tag{4.26}$$

$$F(\omega) = \int_{-\infty}^{\infty} e^{-i\omega x} 2\operatorname{sinc}(2x)\, dx \begin{cases} 0 & \omega < -\frac{1}{2} \\ 2\pi & -\frac{1}{2} < \omega < \frac{1}{2} \\ 0 & \omega > \frac{1}{2} \end{cases} \tag{4.27}$$

It is concluded that Fourier transform of $K(x)$ is Rectangular Function. Hence, $F[K](\omega) = (2\pi)^{-N/2} \int e^{-i\omega x}\ K(x)\ dx\ \geq 0$. Shannon wavelet kernel is an admissible Support Vector Kernel.

4.5.3 Making New Kernels from Kernels

Some of the complicated kernel K_3 can be constructed from existing kernels (say K_1 and K_2). The following propositions have been proven to be valid kernels:

1. $K(x, x') = K_1(x, x') + K_2(x, x')$
2. $K(x, x') = aK_1(x, x')$
3. $K(x, x') = K_1(x, x') \cdot K_2(x, x')$
4. $K(x, x') = f(x)f(x')$
5. $K(x, x') = K_3(\varphi(x), \varphi(x'))$
6. $K(x, x') = \boldsymbol{x^T} A\boldsymbol{x'}$
7. $K(x, x') = \exp(K(x, x'))$

4.5.4 Hybrid Kernel for SVM

Jose et al. developed combinational wavelet kernels by multiplying sinc wavelet kernel with RBF kernel and sinc wavelet kernel with Cauchy kernel [190]. Ramin et al. proposed a novel hybrid kernel by combining thin plate spline RBF kernel and Littlewood-Paley wavelet kernel for least square SVM on structural damage detection [191]. A Linear-Polynomial Kernel was proposed for SVM on speech recognition [192]. Experimental results of these studies show that the hybrid

kernels that were proposed performed better when compared to conventional kernels. A hybrid kernel (SWGK) of Shannon Wavelet Kernel and Gaussian kernel is created and proven to be valid kernel because from Proposition 1, 2 from Section 4.6.3 and the proof of validity of Shannon Wavelet Kernel is proved to be valid kernels.

$$K_{SWG}(x, x') = \prod_{i=1}^{n} (\alpha)\left(2\,\mathrm{sinc}\,2\left(\frac{x_i - x'_i}{a_i}\right) - \mathrm{sinc}\left(\frac{x_i - x'_i}{a_i}\right)\right) + (1+\alpha)(e^{-\gamma\|x-x\prime\|^2}) \quad (4.28)$$

4.5.5 A New Kernel – Champernowne Kernel

A novel kernel is created from Champernowne distribution where the function is used to describe income distribution within a group of people [172][173]. By taking the shape of cosh function, Champernowne distribution is shown as Figure 4.7.

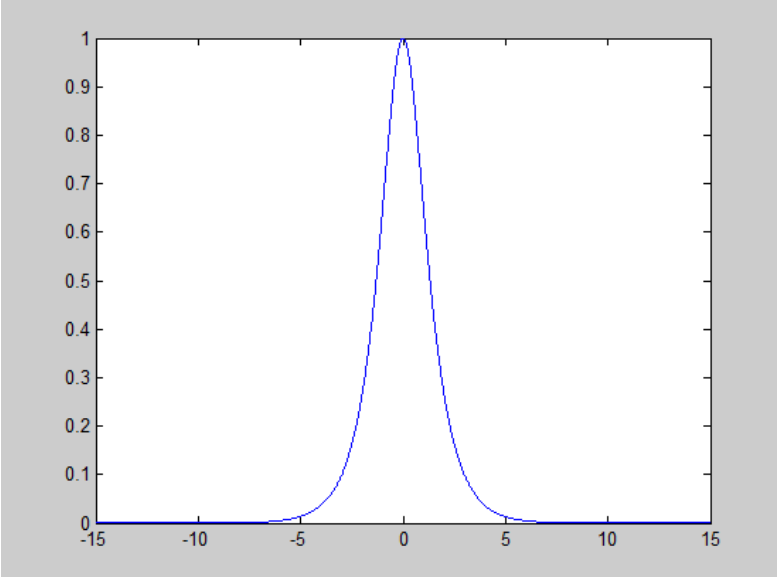


Figure 4.7 Champernowne distribution function

Champernowne distribution can be defined as:

$$f(x) = \frac{n}{\cosh[\alpha(x - x_0)] + \lambda} \tag{4.29}$$

where α, x_0 and λ are positive parameters while n is the normalizing constant.

Normalizing constant in probability function ensures the probability sums up to 1. For kernel, it is not necessary to sum up to 1. Hence, the normalizing constant is omitted [174] [175]. The new kernel is defined as:

$$K_{new}(x, x') = \frac{1}{\cosh[\alpha(x - x')] + \lambda} \tag{4.30}$$

To prove $K_{new}(x, x')$ is valid kernel,

$$\cosh(x) = \frac{e^{x} + e^{-x}}{2} \tag{4.31}$$

From Proposition 1, 2 and 7 in Section 4.6.3, it can be concluded that $K_{new}(x, x')$ is a valid kernel that fulfilled Mercer's Theorem.

In Chapter 5, performance of these newly designed kernels will be used as comparison to the commonly adapted kernels namely Linear, Polynomial and RBF kernels.

4.6 Support Vector Machine Implementation

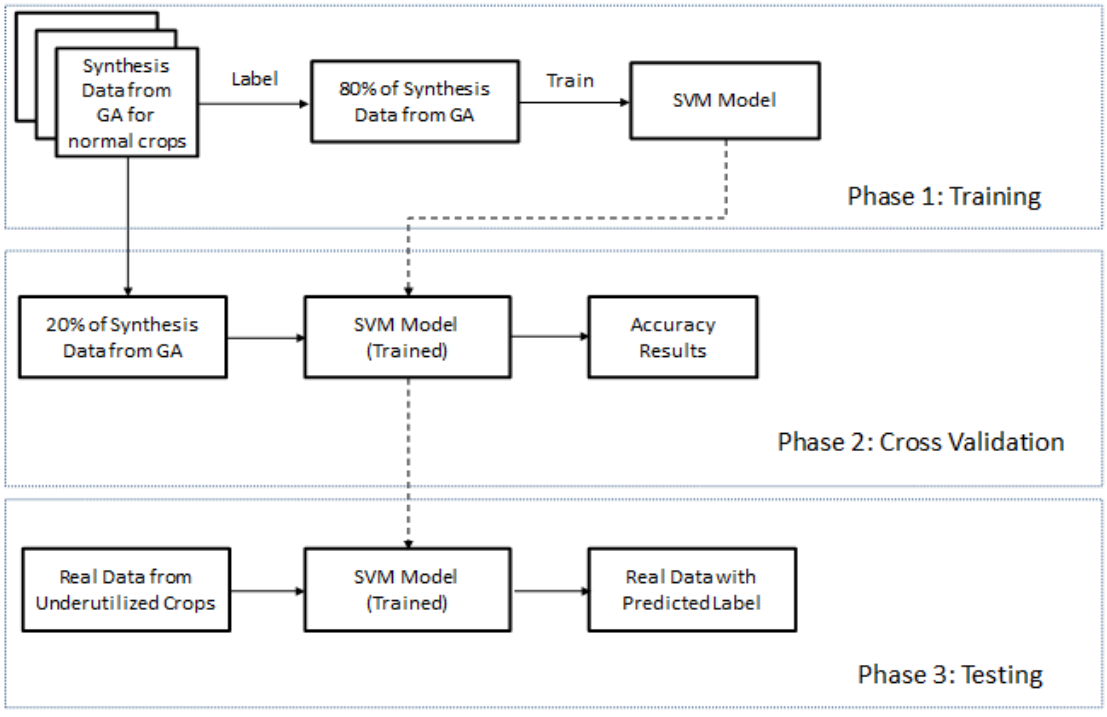


Figure 4.8 SVM Implementation Phases

In Chapter 3, theoretical behind SVM is discussed. In this section, implementation of SVM in this project is presented. Figure 4.8 shows the three phases involved in implementing Support Vector Machine in this project. These three phases are training, cross validation and testing. During the training phase, 80% of the data generated from Genetic Algorithm for normal crops is randomly selected as training data. The other 20% are validating data for cross validation. Training data is labelled by using the technique demonstrated in previous sections. These labelled data is used to train a SVM model. The trained SVM model is then used to cross validate the other 20% data by comparing the data with label predicted by SVM and data that was already labelled. This will give accuracy results. Based on the accuracy, parameters of the model will be tuned to get the best model for real case testing. Real data from underutilised crops are the new test cases to predict the potential of success in economic growth by planting these underutilised crops in the village.

4.7 System Implementation

First stage – Farm Income Models in relation to village characteristics: The relationships among the three main components which are village characteristics

to farm income are to be found. Population, household income, education level are some of the examples that are categorized as village characteristics.

Second Stage –Data Generation by Genetic Algorithm: As mentioned, Genetic Algorithm was used to solve lack of data problem by generating sufficiently large data to train Support Vector Machine. These data were actually values associated with variables of village characteristics.

Third Stage – Data Labelling: Farm income that is greater than target fitness value is labelled as 1; otherwise the set of data will be labelled as 0. Refer to previous section for more details on Data Labelling.

Fourth Stage – Data Classification by Support Vector Machine: At the final stage, real villages will be tested. SVM classify potential of test villages based on characteristics of these villages.

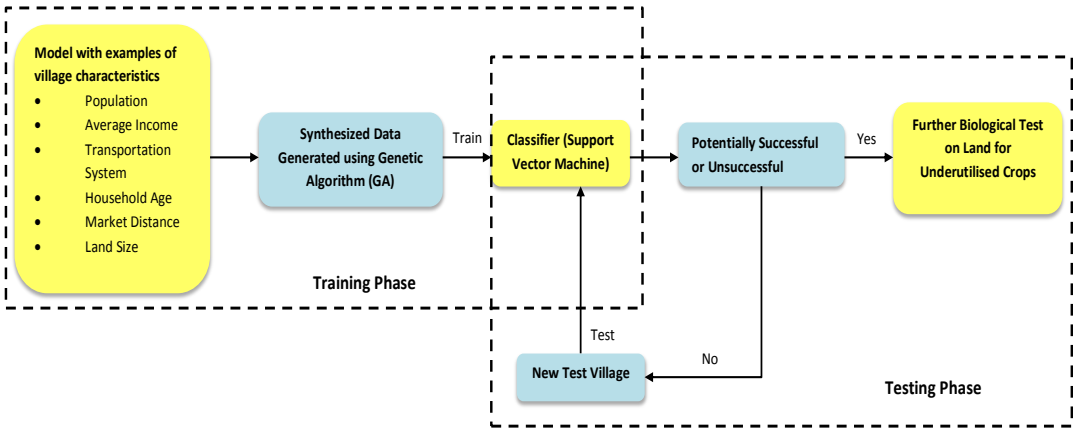


Figure 4.9 Overall System Architecture

4.7.1 Real Application

The block diagram describes the real application of this proposed framework. With this framework, Crops for Future Research Centre is able to identify the required location before actually planting the crops physically in identified village. Farmers will not plant new crops if there is no assurance of successfulness in selling the crops [197].

Chapter 5 Results and Discussions

5.1 Introduction

This section presented the experiment results conducted for this research. Robustness of SVM on approximate models and tested against actual models on different population size. New kernels were tested in following section on benchmark data to test the performance against the commonly used kernels. The new kernels were then added to the other experiments. In order to test for the feasibility on other applications and proximity of synthetic data generated to the real data, model and data from World Bank for GNI per capita were used to implement the proposed framework and tested against other synthetic data generation methods. The proposed framework is then applied on real village data which were provided by Crops For Future Research Centre (CFF). Parameters to adjust the synthetic data produced were also justified.

5.2 Performance Measure

In this chapter, performance of the classifier will be evaluated mainly by using two metrics which are the accuracy and Area under ROC (Receiver Operating Characteristics) Curve (AUC). Accuracy is the ratio between number of correct predictions and total number of data points. ROC curve plots the rate correctly classifying true positive against the rate of incorrectly classifying true negative, as prediction threshold is varied. Area under the curve (AUC) is extracted from ROC curve by taking the value of area under the curve. AUC closer to 1 is

associated with classifier near perfect accuracy when threshold value is correctly defined. If AUC = 0.5, it indicates that the predictions made by the classification model is random. Low AUC which is less than 0.5 means that the classification problem is not correctly set up. This can be fixed by reversing the predictions.

5.3 Robustness of SVM on Approximate Models

5.3.1 Approximate Model 1

In order to prove the robustness of Support Vector Machine (SVM) on approximate models, the following experiment was set up. Two models were involved in this experiment. For each model, population size was set to be 20, 150 and 800 to show the effects of increasing synthetic training data size on SVM classification. The first model was the Ohm's Law in circuit analysis which derived from the relationship between voltage and current in circuit through experimental set up.

Table 5.1 Actual Model vs Approximate Model for Ohm's Law

	Model
Actual Model	$V = IR$
Approximate Model	$V' = \frac{IR}{a} + b - c$

V is the voltage, I is current, R is resistance, a, b and c are the terms added for representation of approximation to the actual model.

The following steps are carried out for this experiment:

1. The actual model in Table 5.1 is used as cost function in Genetic Algorithm
2. Population size is set to 20 in GA. This resulted in 1180 data points.
3. 70% of the data generated in step 2 is used as training data and the rest 30% is used as validation data.
4. Compute the accuracy and AUC for actual model test.
5. Repeat for population size 150 and 800 which resulted in 7650 and 73600 respectively.

Steps 1 to 5 are repeated for approximate model. The only difference is during step 3, the 70% training data comes from approximate model while the validation data comes from the actual model. This is to reflect the use of approximate model in this research to generate artificial data and still be able to achieve acceptable accuracy when classifying actual data.

The test results are presented in Table 5.2, 5.3 and 5.4. Comparison of three standard kernels for SVM (Linear, Polynomial and Radial Basis Function kernel) are also presented.

Table 5.2 Average Accuracy of Experimental Model 1 (Population Size of 20)

Population size: 20				
	Accuracy [AUC]			
	SVM			**BPNN**
	Linear	**Poly**	**RBF**	
Actual Model	94.41 [0.9403]	94.49 [0.9403]	92.46 [0.9403]	-
Approximate Model	69.66 [0.9404]	73.9 [0.9507]	73.9 [0.9688]	94.32

Table 5.3 Average Accuracy of Experimental Model 1 (Population Size of 150)

Population size: 150				
	Accuracy [AUC]			
	SVM			**BPNN**
	Linear	**Poly**	**RBF**	
Actual Model	94.81 [0.9573]	94.82 [0.9573]	94.09 [0.9581]	-
Approximate Model	94.82 [0.9601]	79.19 [0.9846]	79.19 [0.9937]	78.14

Table 5.4 Average Accuracy of Experimental Model 1 (Population Size of 800)

Population size: 800				
	Accuracy [AUC]			
	SVM			**BPNN**
	Linear	**Poly**	**RBF**	
Actual Model	95.80 [0.9710]	95.88 [0.9710]	96.75 [0.9886]	-
Approximate Model	95.84 [0.9719]	82.40 [0.9946]	82.40 [0.9987]	82.40

5.3.2 Approximate Model 2

To further prove that the results are repeatable for more complicated model, experiment steps in Section 5.3.1 are repeated for the following model by replacing cost function in Step 1. This model describes the motion of a moving object at constant acceleration. Different population size in GA is used, which are 20, 150 and 800. These results show different amount of data points generated to train SVM, which are 1020, 7650 and 94400 respectively. The training data is compared between Backpropagation Neural Network (BPNN) and three standard kernels for SVM, which are Linear, Polynomial and Radial Basis Function.

Table 5.5 Actual Model vs Approximate Model for Motion

	Model
Actual Model	$s = ut + \frac{1}{2}at^2$
Approximate Model	$s = \frac{ut + \frac{1}{2}at^2}{c} + b$

s is the distance traveled by the object, u is the velocity of the traveling object, t is the time taken to travel, a is the acceleration of the object, b and c are the terms added to create the inaccuracy in this model.

Table 5.6 Average Accuracy of Experimental Model 2 (Population Size of 20)

Population size: 20				
Condition	**Accuracy [AUC]**			
	SVM			**BPNN**
	Linear	**Poly**	**RBF**	
1	90.39 [0.9012]	91.76 [0.9086]	61.56 [0.9233]	-
2	91.76 [0.8982]	61.37 [0.9277]	66.37 [0.9586]	65.39

Table 5.7 Average Accuracy of Experimental Model 2 (Population Size of 150)

Population size: 150				
Condition	**Accuracy [AUC]**			
	SVM			**BPNN**
	Linear	**Poly**	**RBF**	
1	93.8 [0.9471]	96.99 [0.9794]	96.86 [0.9986]	-
2	94.61 [0.9471]	73.61 [0.9929]	73.63 [0.9996]	73.63

Table 5.8 Average Accuracy of Experimental Model 2 (Population Size of 800)

Population size: 800				
Condition	**Accuracy [AUC]**			
	SVM			**BPNN**
	Linear	**Poly**	**RBF**	
1	97.07 [0.9837]	98.06 [0.9900]	98.20 [0.9947]	-
2	97.08 [0.9852]	91.51 [0.9935]	91.51 [0.9969]	78.62

5.3.3 Section Analysis

From the Table 5.6, 5.7 and 5.8, it can be seen that as the amount of training data increases, the training of SVM gets better across all the kernels. As expected, regardless of the complexity of the model, performance of SVM is consistent, with the accuracy of the classifier getting better when the amount of training data increases across all the kernels. Linear kernel performs the best among the kernels. From this set of experiment, it can be concluded that SVM trained by data generated from approximate model is able to classify data from actual model. The following table illustrates the relationship between the models from this experiment and the economic models which are used in this research.

Table 5.9 Relationship between Models

	Experiment 1	**Experiment 2**	**Economic Model**
Output from actual model	V	s	Farmer income from exact model
Output from approximate model	V'	s'	Farmer income from inaccurate model
Variables	I	u, a	Parameters of inaccurate model
Noise Terms	a, b, c	a, b, c	-

Actual Model	$V = IR$	$s = ut + \frac{1}{2}at^2$	Exact model unknown
Approximate Model	$V' = \frac{IR}{a} + b - c$	$s' = \frac{ut + \frac{1}{2}at^2}{c} + b$	See Model 1, 2, and 3 in Section 4.3.1.

5.4 Comparison of new kernels on benchmark data

In this section, 5 sets of benchmark data were used to compare the classification performance of newly designed kernels which have been validated as valid kernels in previous chapter with the commonly used kernels. The commonly used kernels are linear, polynomial and RBF. Three new kernels namely Shannon Wavelet, Hybrid SWG and Champernowne are added. Details of the benchmark data were described in Table 5.10.

Table 5.10 Benchmark Data Characteristics

Data	**Data Name**	**Total Data (Training Data/Test Data)**	**Features**
Data 1	Heart	270 (150/120)	13
Data 2	Diabetes	768 (461/307)	8
Data 3	Australian	690 (414/276)	14
Data 4	German.numer	1000 (600/400)	24
Data 5	ionosphere	351 (211/140)	34

Table 5.11 Performance Comparison of between Commonly Used Kernels and New Kernels

Kernels	Data 1	Data 2	Data 3	Data 4	Data 5	Average Accuracy
Linear	85	79.15	83.70	76.5	93.57	**83.58**
Polynomial	82.5	69.71	84.06	73.5	78.57	77.67
RBF	83.33	82.09	83.70	75	97.86	**84.4**
Shannon Wavelet	82.5	79.80	84.06	74.25	91.43	82.41
Hybrid SWG	83.33	79.48	84.06	74.25	91.43	82.51
Champernowne	83.33	80.13	83.70	78.5	96.43	**84.42**

5.4.1 Section Analysis

As can be observed from Table 5.11, the performance of the new kernels is as comparable as the widely used kernels namely RBF and linear kernels depending on the data. Champernowne achieved the best average accuracy on the benchmark data. The performance of hybrid kernel - Hybrid SWG is better than Polynomial kernel. For the following experiments, the new kernels will be added alongside with commonly used kernels to further examine these new kernels.

5.5 Feasibility on other applications and proximity of synthetic data generated to the real data

In this section of the experiment, dataset downloaded from World Bank is used. The purpose of this experiment is to demonstrate the application of the

methodology on real data that suffer from insufficiency. The mathematical model to calculate Gross National Income (GNI) per capita is taken directly from the World Bank website [198]. Gross National Income is the measure of country's income and is defined as the sum of all income earned, including from overseas, by a country. Details of the methodology by using Atlas method can be found on the website. Atlas method is a conversion factor that is used to reduce volatility effect of exchange rates across different countries in comparing national income. The following model is used as fitness function for GA to generate artificial training data as detailed in previous chapter:

$$r_{t-n} = \frac{p_t}{p_{t-n}} \tag{5.1}$$

$$r_{t-n}{}^{SDR\$} = \frac{p_t{}^{SDR\$}}{p_{t-n}{}^{SDR\$}} \tag{5.2}$$

$$e_t{}^{atlas} = \frac{1}{3}\left[e_t + e_{t-1}\left(\frac{r_{t-1}}{r_{t-1}{}^{SDR\$}}\right) + e_{t-2}\left(\frac{r_{t-2}}{r_{t-2}{}^{SDR\$}}\right)\right] \tag{5.3}$$

$$Y_t{}^{atlas\$} = Y_t / e_t{}^{atlas} \tag{5.4}$$

$$Y^{per\ capita} = Y_t{}^{atlas\$} / N_t \tag{5.5}$$

where $e_t{}^{atlas}$ refers to Atlas method exchange rate, r_{t-n} is inflation rate of the country between year t and year $t-n$, p_t is GDP deflator of the country,

$r_{t-n}{}^{SDR\$}$ is the international inflation rate between year t and year $t-n$, $p_{t-n}{}^{SDR\$}$ is the Special Drawing Rights deflator (details of SDR deflator is described in details in [199]), e_t is average annual exchange rate for year t, $Y_t{}^{atlas\$}$ is the GNI in U.S. dollars after conversion from local currency and $Y^{per\ capita}$ is GNI per capita. The data is summarised as the following plots:

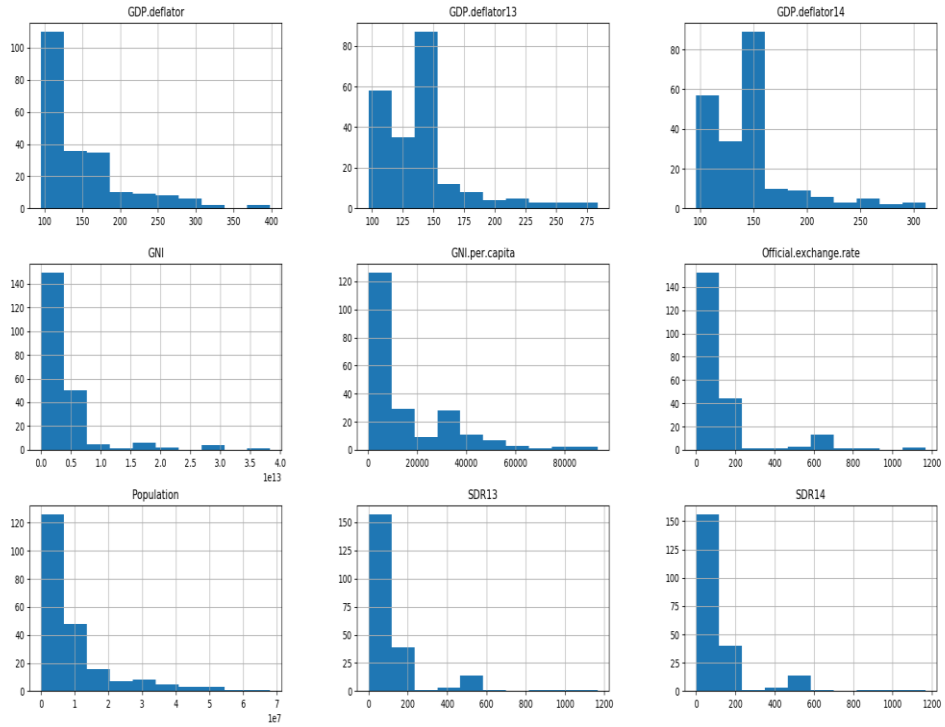


Figure 5.1 Data Summary Plots of Variables for GNI Model

The training data generated is labelled according to the World Bank classification on low income and high income countries [200]. It is classified as low-income economies for those countries with a GNI per capita, calculated

using the World Bank Atlas method, of $995 or less; lower middle-income economies are those with a GNI per capita between $996 and $3,895; upper middle-income economies are those with a GNI per capita between $3,896 and $12,055; high-income economies are those with a GNI per capita of $12,056 or more. Countries with low income and lower middle income economies are labelled as 0 and countries with upper middle income and high income are classified as 1.

5.5.1 World Bank GNI Data with Synthetic Data Training and Test

After generation of data, the data is split into 70% training data and remaining 30% as validation data set. The training data is standardized through feature scaling to reduce the impact of large values on classifier from some of the variables in the dataset. Data from 218 countries are then used as test data to validate the model. The results are presented in the following table.

Table 5.12 Average Accuracy of GNI Model

		Accuracy [AUC]	
		Validation	**Test**
SVM	**Linear**	98.30 [0.9996]	96.33 [0.9996]
	Poly	96.92	93.12

		[0.9931]	[0.9931]
	RBF	98.28 [0.9936]	97.71 [0.9936]
	Shannon	93.33 [0.9991]	91.74 [0.9991]
	SWG	93.88 [0.9988]	93.12 [0.9988]
	Champernowne	94.06 [0.9864]	94.95 [0.9864]
BPNN		-	97.71

5.5.2 Comparison of Proposed Methodology with Other Methods

One of the common ways to generate synthetic data is through the Monte Carlo method. This is done by observing the statistic distributions of the data and generates large number of data randomly according to the distribution. [62] [201]. As such, the Monte Carlo Method of generating synthetic data is used to compare with the proposed methodology.

While there are many varieties of Monte Carlo method, the standard process of the method involved defining the parameters for each feature from the data identified. Based on the parameters defined, data is generated following the distribution of the parameters.

In this section, two methods of data generation are used to compare with the proposed methodology. The first method is data generation based on Normal (Gaussian) distribution. From the 218 datasets that were downloaded from World Bank website [198], the mean μ and standard deviation σ of the data were calculated. These parameters were used to generate synthetic data that follows Gaussian distribution. The synthetic data generated is then used as training data to train classifiers. The results that were obtained are recorded as follows:

Table 5.13 Gaussian Distribution Synthetic Training Data Generation Performance

		Accuracy [AUC]	
		Validation	**Test**
SVM	**Linear**	88.05 [0.5031]	36.70 [0.5031]
	Poly	88.05 [0.5255]	36.70 [0.5255]
	RBF	88.06 [0.7145]	36.70 [0.7145]
	Shannon	87.33 [0.5583]	37.61 [0.5583]
	SWG	87.33	37.61

		[0.9461]	[0.9461]
	Champernowne	88.79 [0.5690]	36.70 [0.5690]
BPNN		-	36.70

In this section, minimum and maximum values of each attribute were obtained from the data available. Synthetic data is generated within the range of these minimum and maximum values following uniform distribution. The results that were obtained as follows:

Table 5.14 Uniform Distribution Synthetic Training Data Generation Performance

		Accuracy [AUC]	
		Validation	**Test**
SVM	**Linear**	96.1 [0.5213]	36.70 [0.5213]
	Poly	96.1 [0.5252]	36.70 [0.5252]
	RBF	96.1 [0.8925]	36.70 [0.8925]
	Shannon	96.21	36.70

		[0.6183]	[0.6183]
	SWG	96.48 [0.9962]	36.70 [0.9962]
	Champernowne	96.48 [0.5537]	36.70 [0.5537]
BPNN		-	36.70

In this section, methodology proposed in [28] is replicated on this dataset as comparison with proposed method. The results are presented as follows:

Table 5.15 Synthetic Training Data Generation Proposed by Yang et al. [28] *Performance*

		Accuracy [AUC]	
		Validation	**Test**
SVM	**Linear**	93.58 [0.9759]	92.20 [0.9759]
	Poly	89.79 [0.9564]	89.91 [0.9564]
	RBF	95.41 [0.9861]	93.58 [0.9861]
	Shannon	94.95	86.70

		[0.9853]	[0.9853]
	SWG	95.30 [0.9899]	85.32 [0.9899]
	Champernowne	92.32 [0.9775]	90.37 [0.9775]
BPNN		-	93.58

5.5.3 Section Analysis

As comparison between the method proposed by Yang et al. [28], two data generation methods and proposed methodology, the table as follows are presented where Average Accuracy and Average AUC is calculated by averaging the accuracy and AUC obtained using three different kernels in SVM.

Table 5.16 Compare between proposed methodology and other methods

Methodology	**SVM**	
	Average Accuracy (%)	**Average AUC**
Uniform distribution generator	36.70	0.6361
Gaussian distribution generator	36.70	0.6845
Methods proposed by Yang et al. [28]	89.68	0.9785

Proposed methodology	**94.50**	**0.9951**

As observed from the Table 5.16, the accuracy of proposed method is greatest among all the other methods. AUC shows that the data generator based on Gaussian distribution and Uniform distribution is almost randomly classifying the data. This is due to the data generated is not based on relationships between the features but merely distribution from the parameters. Hence, a data generator based on merely data distribution function is not reflective of real data. From Table 5.16, the overall performance of the proposed methodology proves that data generated from the model is feasible and proximate to the real data which is why the data can be correctly classified and the AUC shows that the data is not randomly classified.

5.6 Parameters Adjustment in Genetic Algorithm

In Genetic Algorithm, there are two parameters that can be adjusted to obtain desirable set of data to be generated. These two parameters are population size and initial population range. Population size determines the number of chromosomes in each generation. Small population size causes GA to operate only in small search space which leads to local optimization instead of global optimization. However, large population size leads to slow computation time. In our case, population size determines the size of the artificial data to be generated.

The following section justifies the number of population sizes for the models in order to generate sufficient data without compromising computational time. Initial population range specifies the range of individuals in initial population. This helps GA to explore wider range of search space to avoid local optimization by specifying the range where GA should also search. In the proposed method, by specifying the initial population, it avoids converging too fast to local optima point. Converging quickly causes imbalanced data being generated for the two classes labelling. Hence, specifying initial population range helps to generate balanced data to label the two classes. Details will be discussed later in the following section.

5.6.1 Parameter 1: Population Size

The three models selected in Chapter 4 are used. Population size of 10, 100, 1000 and 10000 are used to determine the optimal population size to use in terms of generating a large data set without comprimising computational time.

Model 1

Table 5.17 Population Size vs Elapsed Time of Model 1

Population Size	**Elapsed Time (seconds)**	**Dimension of Generated Data**
10	0.782916	510x5
100	0.972541	5100x5

1000	2.890856	51000x5
10000	21.072679	510000x5

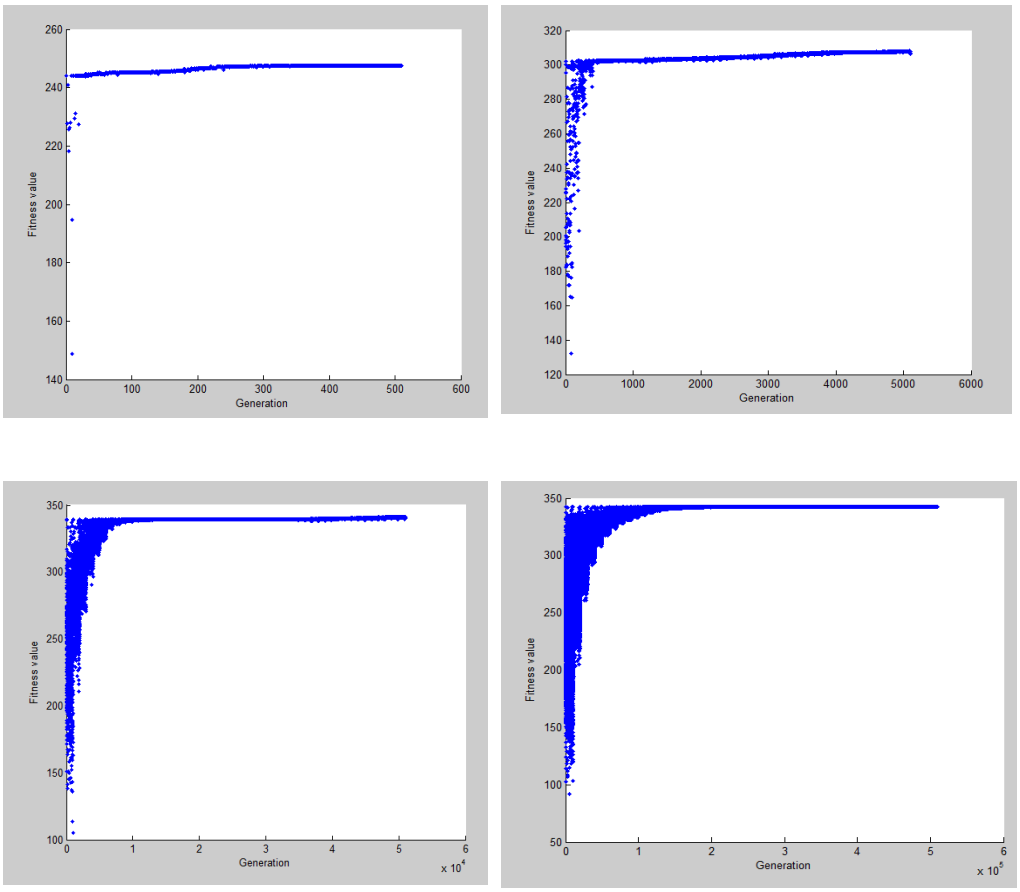


Figure 5.2 Generation versus Fitness Value Plots of Model 1 for Population Size 10 (Top Left), 100 (Top Right), 1000 (Bottom Left) and 10000 (Bottom Right)

Model 2

Table 5.18 Population Size vs Elapsed Time of Model 2

Population Size	**Elapsed Time (seconds)**	**Dimension of Generated Data**
10	0.777262	510x11
100	1.013127	5100x11
1000	3.320471	51000x11
10000	24.659098	510000x11

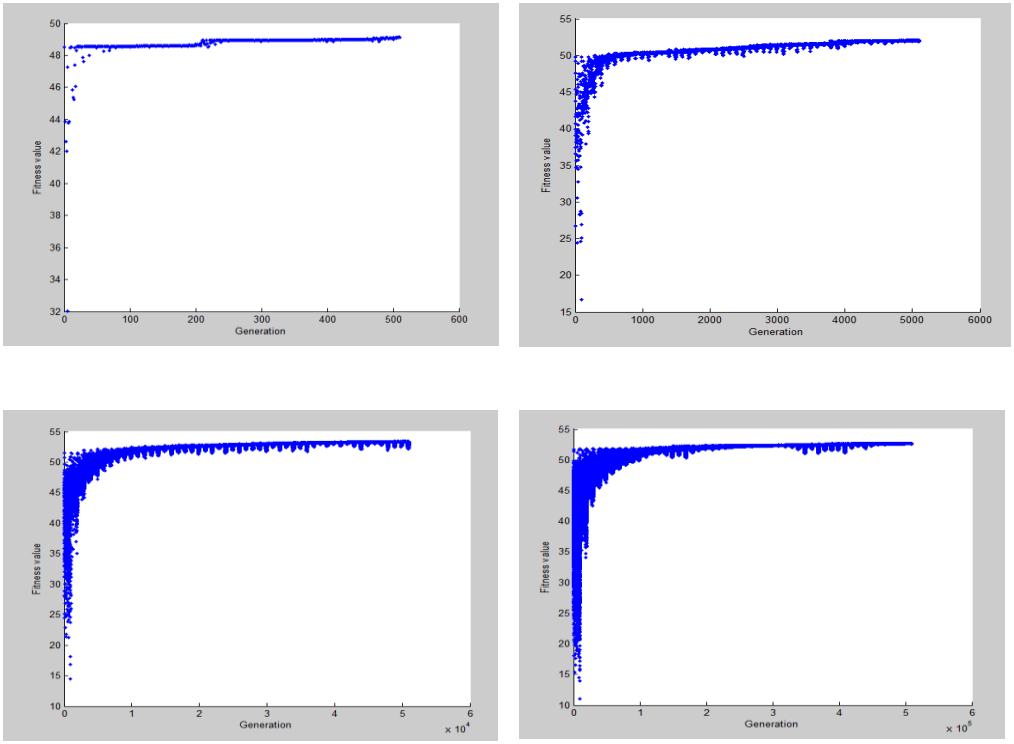


Figure 5.3 Generation versus Fitness Value Plots of Model 2 for Population Size 10 (Top Left), 100 (Top Right), 1000 (Bottom Left) and 10000 (Bottom Right)

Model 3

Table 5.19 Population Size vs Elapsed Time of Model 3

Population Size	**Elapsed Time (seconds)**	**Dimension of Generated Data**
10	0.798411	510x7
100	0.936447	5100x7
1000	2.778618	51000x7
10000	21.677956	510000x7

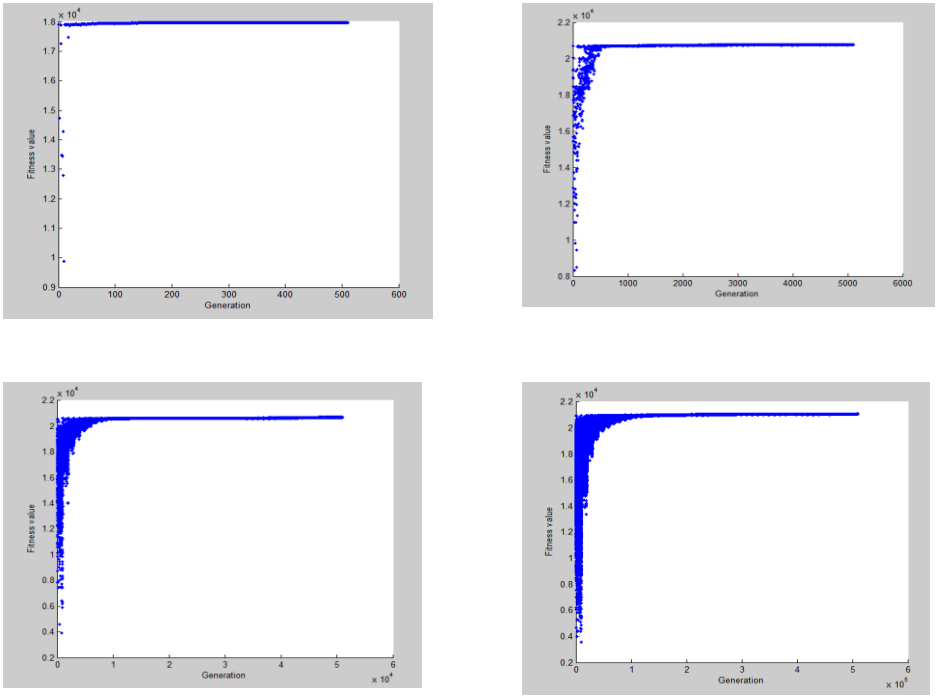


Figure 5.4 Generation versus Fitness Value Plots of Model 3 for Population Size 10 (Top Left), 100 (Top Right), 1000 (Bottom Left) and 10000 (Bottom Right)

Figure 5.1, Figure 5.2 and Figure 5.3 were plotted according to Table 5.1717, Table 5.18 and

Table 5.195.19. From the tables and plots, it can be observed that the computational time increased steeply when population size increases from 1000 to 10000. Therefore, it is suggested to use population size of 1000 in order to generate sufficient data, thereby making computation time reasonable. This can be observed from the figure plotted in Figure 5.4.

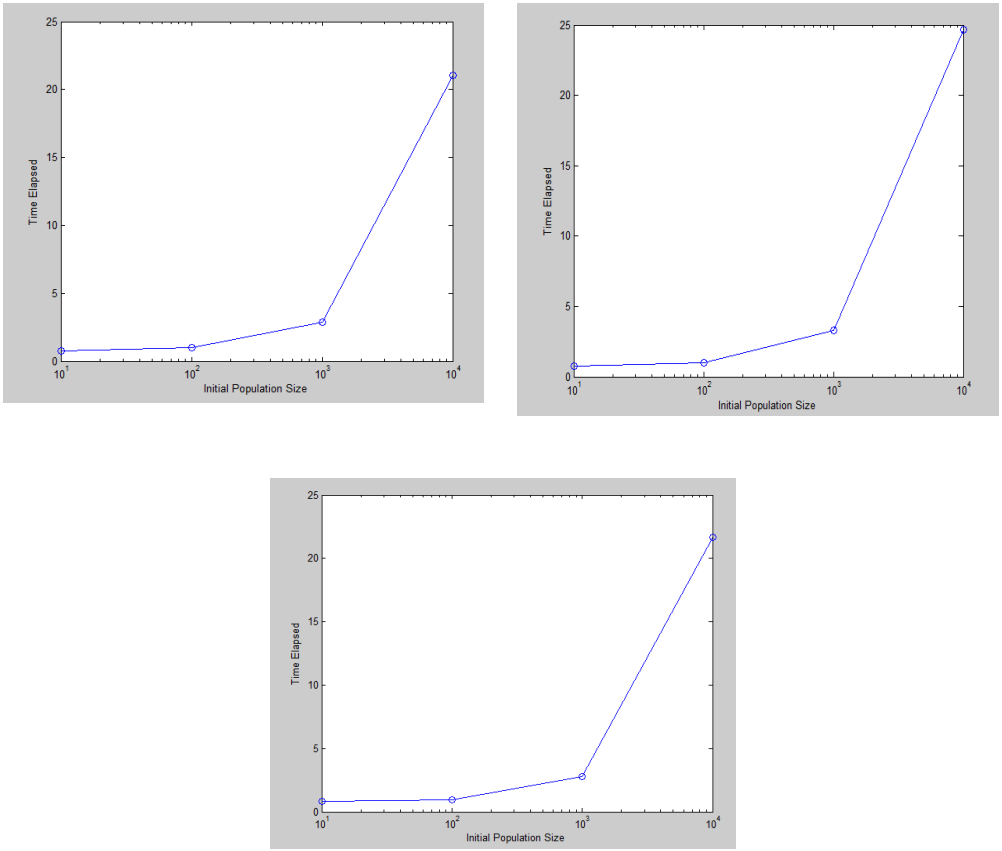


Figure 5.5 Population Size vs Elapsed Time Plots for Model 1(Top Left), 2(Top Right) and 3 (Bottom)

5.6.2 Parameter 2: Initial Population Range

As mentioned earlier, by specifying the initial population, GA is able to explore search space and converge to global optimum point which helps to generate balanced data to label the two classes. Figures below show that the differences when different initial population range are specified. It can be observed that by specifying certain initial population range, the generated data is balanced as desired. By comparing Figure 5.5 to Figure 5.6, the former has two classes which are well balanced, while the latter biased to only one class. The data points within red circle are to be labelled in one class and the rest in another class.

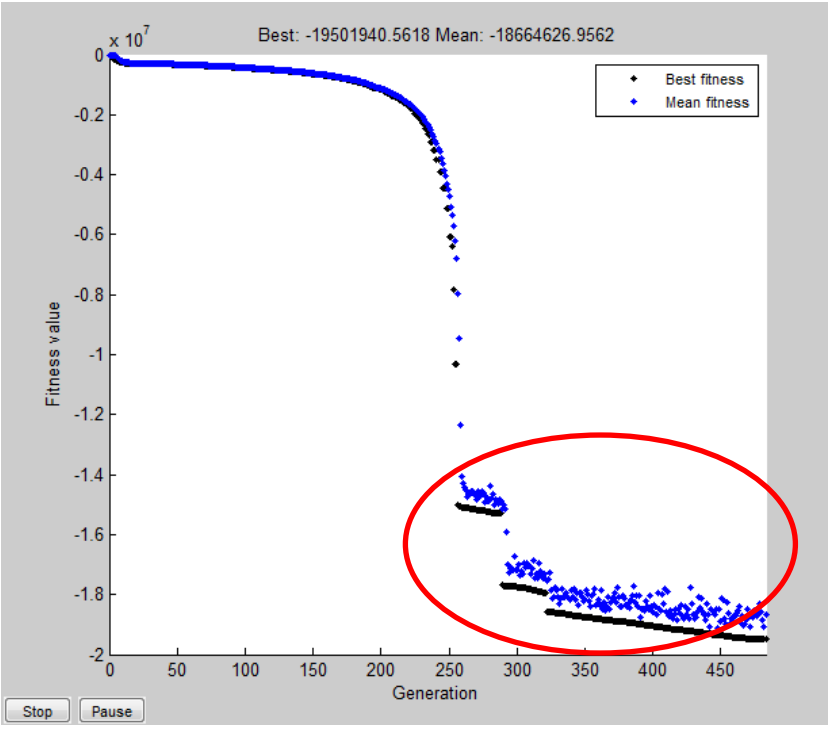


Figure 5.6 Balanced Example from Generation versus Fitness Value Plot

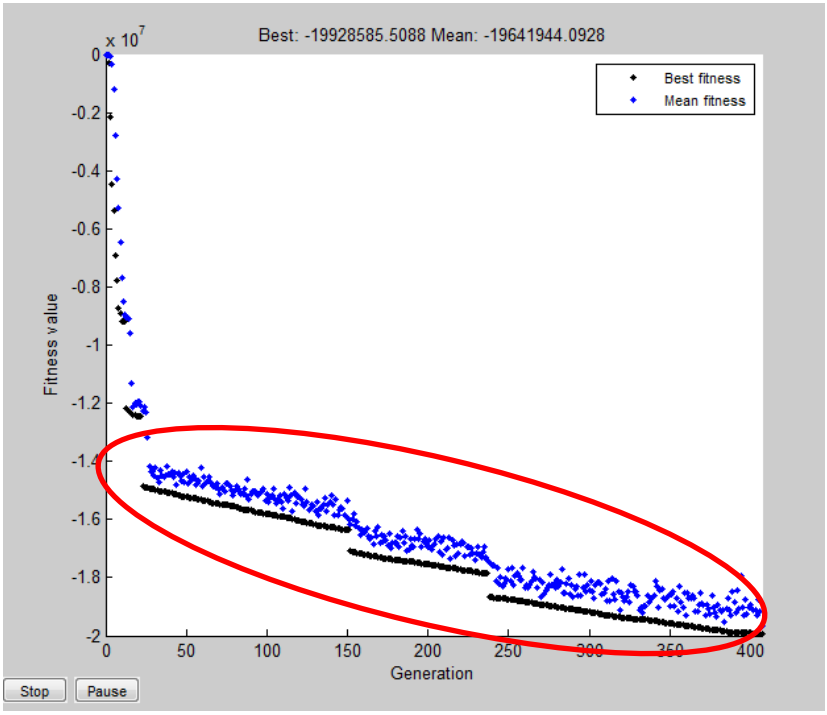


Figure 5.7 Imbalanced Example from Generation versus Fitness Value Plot

5.7 Real Village Data Validation

In this experiment, the three models which were selected in Chapter 4. A set of real data on underutilised crops from 30 villages in Sri Lanka were collected. The data was obtained through Crops for Future Research Centre. This set of data was used to validate the methodology of generating artificial data from approximate model to train SVM, which then will be used to classify potential of successfulness in commercializing underutilised crops. This has been described in details in Chapter 4. The results are presented in the following for Model 2.

Table 5.20 Average Accuracy of Model 2 on Real Village Data

		Accuracy [AUC]	
		Validation	**Test**
SVM	**Linear**	98.33 [0.9977]	70 [0.9977]
	Poly	98.32 [0.9972]	76.67 [0.9972]
	RBF	98.43 [0.9971]	70 [0.9971]
	Shannon	98.68 [0.9946]	70 [0.9946]
	SWG	98.68 [0.9955]	70 [0.9955]
	Champernowne	98.98 [0.9689]	70 [0.9689]
BPNN		-	66.67

Model 2 has been chosen to generate artificial data to train SVM because the variables of real data from Sri Lanka collected match most of the variables in

Model 2. Table 5.21 is presented to provide a clearer view of variables involved in the data and the three models.

Table 5.21 Comparison of Variables Between Real Village Data and Model 1,2 and 3

Real Village Data	**Model 1**		**Model 2**		**Model 3**	
Family Member	Cultivated rice land value		Age	✓	Household Age	✓
Gender	Total family labor inputs		Gender	✓	Household Size	✓
Education	Total material inputs such as seeds, pesticide and fertilizer		Marital Status		Labor	
Age	Farmer's ability index (Education, Experience, Knowledge)		Formal Education	✓	Value of assets	

Number of years involve in farming	Damaged rice acreage (% to total planted acreage)		Experience	✓	Market Distance	
Land size			Extension (Meeting with Extension Agent)		Operational Area	✓
Land ownership			Distance from house to farm	✓	Weather Condition	✓
Distance from house to farm			Household Size	✓		
Climatic factors			Secondary Occupation			
			Land Status	✓		
			Land Size	✓		

Besides the Area under ROC curve to ensure the classification is not random, the predicted label was also manually compared with the farm income from village in Sri Lanka. The income is higher for village 1 to village 10 as compared to the other villages in the real data set. Hence, it was expected that the SVM classifies first 10 villages to be class 1 while the rest class -1. As can be seen from the Figure 5.7, the correctly classified labels in class 1 were among the 10 villages. The other class 1 villages were misclassified due to the inaccuracy of the model of normal crops on underutilised crops.

predict_label <30x1 double>

	1	2
1	-1	
2	1	
3	-1	
4	1	
5	-1	
6	-1	
7	-1	
8	1	
9	-1	
10	-1	
11	-1	
12	-1	
13	-1	
14	-1	
15	-1	
16	-1	
17	-1	
18	-1	
19	-1	
20	-1	
21	-1	
22	-1	
23	-1	
24	-1	
25	-1	
26	-1	
27	-1	
28	-1	
29	-1	
30	-1	

Figure 5.8 Predicted Labels of SVM Trained by Data Generated from Model 2

Model 3

The results for Model 3 are presented in the table below. Although classification accuracy shows that the model is able to correctly classify 20 out of 30 villages, however, AUC shows that the classification model is not performing better than random guessing the label. This opinion is also further justified by observing the predicted label from SVM being biased towards class -1 as can be seen from Figure 5.8. Hence, it is deduced that Model 3 is not suitable for this set of data obtained from Sri Lanka villages as the variables involved are not providing information to train SVM.

Table 5.22 Average Accuracy of Model 3 on Real Village Data

		Accuracy [AUC]	
		Validation	**Test**
SVM	**Linear**	77.35 [0.4988]	66.67 [0.4988]
	Poly	77.35 [0.4988]	66.67 [0.4988]
	RBF	77.35 [0.4988]	66.67 [0.4988]
	Shannon	95.06	33.33

		[0.4384]	[0.4384]
	SWG	95.06 [0.4383]	33.33 [0.4383]
	Champernowne	95.06 [0.4384]	33.33 [0.4384]
BPNN		-	66.67

predict_label <30x1 double>

	1	2
1	-1	
2	-1	
3	-1	
4	-1	
5	-1	
6	-1	
7	-1	
8	-1	
9	-1	
10	-1	
11	-1	
12	-1	
13	-1	
14	-1	
15	-1	
16	-1	
17	-1	
18	-1	
19	-1	
20	-1	
21	-1	
22	-1	
23	-1	
24	-1	
25	-1	
26	-1	
27	-1	
28	-1	
29	-1	
30	-1	

Figure 5.9 Predicted Labels of SVM Trained by Data Generated from Model 3

5.8 Summary

While the classification performance for the proposed methodology is proven through various experiments, the same approach has been applied on regression problem of predicting the cost of underutilised crops based on models and data obtained from normal cash crops. The details and results are provided in Appendix A of this thesis.

Chapter 6 Conclusion and Future Works

6.1 Conclusion

This research is motivated by solving the issue, absence of data on underutilised crops to predict potential of economic success by commercialising underutilised crop in rural villages. This is solved by generating artificial data using approximate models as objective function of Genetic Algorithm. The data generated is used to train machine learning algorithm, namely Support Vector Machine to mitigate the inaccuracies induced by approximate models and formulation of model itself.

The main aim of this research is to develop a hybrid Genetic Algorithm based approximate cash crop model with Support Vector Machine classifier framework that can be used to predict economic viability of underutilised crop in rural area. This is done by proving experimentally that it is possible to achieve good classification accuracy by training the SVM with data generated from an approximate model used as the fitness function in a Genetic Algorithm. The Genetic Algorithm were used to generate the artificial data because real world constraints on the physical characteristics of interest to us for rural villages are easily included in the methodology. It is one of the aims of this research to show that the data generated artificially has a high probability of corresponding to real villages since we include real life constraints in the algorithm and hence this

artificial data can then legitimately be used to train the Support Vector Machine. In certain domains "big data" is not available. The machine learning techniques are extended into this domain by proposing a methodology with which to generate training data by running a genetic algorithm which produces sets of parameters of different fitness. In this research, it is not the fittest solution which is important but a set of fittest solutions which give the opportunity to look in the real world for corresponding candidates (which are the rural villages in this research) which we propose to relevant authorities as "most likely to succeed" in commercialising underutilised crops. The SVM plays the role in processing "test data" obtained from real world locations and then classifying those locations as likely to succeed or otherwise. Soft margin SVM allows minimization of trade off between training error and margin separation. In addition, cost of misclassification parameter, C controls the margin of separation to avoid overfitting.

Kernel plays a crucial role in SVM by implicitly mapping the input space of nonlinear data to a higher dimensional feature space where data can be linearly separable. In this research, three commonly used kernels (Linear, Polynomial and RBF) are adapted. Three other kernels (Shannon Wavelet, Hybrid Shannon Wavelet Kernel and Gaussian kernel, Champernowne) are introduced. Champernowne distribution is generally used to describe income distribution of

a group of people. This makes it ideal for data generated from farm income models which are used as fitness function in Genetic Algorithm. These kernels are proved to be a valid kernel through Mercer's Theorem and tested against selected benchmark data before implementing to the proposed framework for further testing. The purpose of these kernels is to improve the performance of the framework. Champernowne kernel is the best performer based on classification accuracy on benchmark data as compared to other kernels. It is also as comparable as other popular kernels on proposed framework.

Finally, the results show that the capability of correctly identifying villages which have the potential of successfully growing underutilised crops by training the SVM with artificial data generated from a model which was developed for normal cash crops. More villages are intended to be tested as more underutilised crops are commercialised in rural areas.

6.2 Future Works

The experimental results have demonstrated the effectiveness of the proposed methodology in providing data to train SVM from equivalent model. The proposed methodology can be further improved and developed.

The proposed methodology has been applied on classification problem where villages are classified based on successfulness of implementing underutilised

crops. This can be extended to regression problem where Support Vector Regression (SVR) can be used to predict the value of farm income. A report on using the same approach for another regression problem is provided in Appendix A. In this report, the regression problem is attempted to predict the cost of underutilised crop (Bambara groundnut). This is done by training SVR with artificial data generated from GA where equivalent normal cash crop (peanut) model were used as fitness function. However, the lack of underutilised crops data is replaced with data from other crops. These preliminary experiments of applying the methodology on regression problem could be improved when underutilised crop data is available.

Another direction of improvement could be from replacing SVM with other classifiers such as decision tree, random forest and so on. The results could be used as comparative analysis. This can also be used to improve the system for user to be able to choose classifier from the list to predict the successfulness of the village which they input into the system.

In order to further validate the effectiveness of the methodology, more real data on villages that planted underutilised crops successfully and unsuccessfully can be acquired for testing. Furthermore, the methodology can be tested in other applications that are facing the problem which data and labels are absent.

References

[1] "Enhancing the contribution of neglected and underutilized species to food security , and to incomes of the rural poor," Rome, 2000.

[2] P. Chivenge, T. Mabhaudhi, A. T. Modi, and P. Mafongoya, "The potential role of neglected and underutilised crop species as future crops under water scarce conditions in Sub-Saharan Africa," *Int. J. Environ. Res. Public Health*, vol. 12, no. 6, pp. 5685–5711, 2015.

[3] T. Mabhaudhi, P. O'Reilly, S. Walker, and S. Mwale, "Opportunities for underutilised crops in Southern Africa's post-2015 development agenda," *Sustain.*, vol. 8, no. 4, pp. 1–16, 2016.

[4] "Growing bambara groundnut in Malaysia," *Crops For the Future*. [Online]. Available: http://www.cffresearch.org/Updates-@-Growing_bambara_groundnut_in_Malaysia.aspx#sthash.bePtxlt3.dpbs. [Accessed: 01-May-2017].

[5] "Improvement of post-harvest and enhancement of fonio in Africa (Aval Fonio)," *Aval Fonio*. [Online]. Available: https://aval-fonio.cirad.fr/en. [Accessed: 01-May-2017].

[6] I. Jovanovic and J. Enright, "An approximate model for pulsar navigation simulation," *Acta Astronaut.*, vol. 119, pp. 101–109, 2016.

[7] A. J. Bahr and K. R. Clausing, “An Approximate Model for Artificial Chiral Material,” *IEEE Trans. Antennas Propag.*, vol. 42, no. 12, pp. 1592–1599, 1994.

[8] M. Hofer, M. Muehlebach, and R. D. Andrea, “Application of an Approximate Model Predictive Control Scheme on an Unmanned Aerial Vehicle,” pp. 2952–2957, 2016.

[9] A. Fujimoto, “An Economic Analysis of Peasant Rice Fanning in Kelantan, Malaysia,” *South East Asian Stud.*, vol. 14, no. 2, pp. 159–176, 1976.

[10] K. Assis, Z. Nurul Azzah, and A. Mohammad Amizi, “RELATIONSHIP BETWEEN SOCIOECONOMIC FACTORS , INCOME AND PRODUCTIVITY OF FARMERS : A CASE STUDY ON PINEAPPLE FARMERS,” *Int. J. Res. Humanit. Arts Lit.*, vol. 2, no. 12, pp. 67–78, 2014.

[11] T. A. Hassan, “Economic Analysis of Factors Affecting the Farmer Income Under Traditional Farming System in South Darfur State – Sudan 2 . Economic Activities in South,” vol. 1, no. 3, pp. 114–119, 2015.

[12] A. Karunaratne, S. N. Azam-ali, S. Walker, and A. C. Ruane, “Modelling

the productivity of underutilised crops for climate resilience," no. September 2015, 2017.

[13] A. S. Karunaratne, S. Walker, and S. N. Azam-Ali, "Assessing the productivity and resource-use efficiency of underutilised crops: Towards an integrative system," *Agric. Water Manag.*, vol. 147, pp. 129–134, 2015.

[14] T. Mabhaudhi, "Status of Underutilised Crops in South Africa : Opportunities for Developing Research Capacity," 2017.

[15] "CropBASE."

[16] A. O. Chukwukere, J. Sulaiman, and A. Chidi Nwosu, "Farm – Level Determinants of Agricultural Commercialization," *Int. J. Agric. For.*, vol. 2, no. 2, pp. 1–5, 2012.

[17] I. Bobojonov and A. Aw-Hassan, "Impacts of climate change on farm income security in Central Asia: An integrated modeling approach," *Agric. Ecosyst. Environ.*, vol. 188, pp. 245–255, 2014.

[18] A. W. Ebert, "Potential of Underutilized Traditional Vegetables and Legume Crops to Contribute to Food and Nutritional Security, Income and More Sustainable Production Systems," *sustainability*, vol. 6, pp.

319–335, 2014.

[19] W. Savage, “A walk through the commercialization process,” *Missouri Business*. [Online]. Available: https://missouribusiness.net/article/commercialization-process/. [Accessed: 01-Jul-2017].

[20] J. Von Braun, H. De Haen, and J. Blanken, “Research Report of Agriculture Under Population Pressure : Effects on Production , Consumption , and Nutrition in Rwanda,” 1991.

[21] “Social and economic issues,” *FAO Corporate Document Respository*. [Online]. Available: http://www.fao.org/docrep/v9926e/v9926e08.htm. [Accessed: 01-Jul-2017].

[22] G. G. Daniela Horna, Samuel Timpo, “Marketing Underutilized Crops: The case of the African garden egg (Solanum aethiopicum) in Ghana,” Rome, 2007.

[23] E. Gotor and F. Caracciolo, “Improving rural livelihoods through the conservation and use of underutilized species : Evidence from a community research ...,” *Int. J. Agric. Sustain.*, vol. 11, no. 4, pp. 347–362, 2013.

[24] S. B. Ravi, T. K. Hrideek, A. T. K. KumarMal, T. R. Prabhakaran, B. Mal, and S. Padulosi, “Mobilizing neglected and underutilized crops to strengthen food security and alleviate poverty in india,” *Indian J. Plany Genet. Resour.*, vol. 23, no. 1, pp. 110–116, 2010.

[25] G. Gruère, A. Giuliani, and M. Smale, “Marketing Underutilized Plant Species for the Benefit of the Poor : A Conceptual Framework,” Washington, 2006.

[26] Q. H. Fei Hu, Ed., *Intelligent Sensor Networks*. CRC Press, 2016.

[27] P. Niyogi, F. Girosi, T. Poggio, and A. Member, “Incorporating Prior Information in Machine Learning by Creating Virtual Examples,” vol. 86, no. 11, 1998.

[28] J. Yang, X. Yu, Z.-Q. Xie, and J.-P. Zhang, “A novel virtual sample generation method based on Gaussian distribution,” *Knowledge-Based Syst.*, vol. 24, no. 6, pp. 740–748, 2011.

[29] H. Hähnel, A.-J. Hempel, and G. Herbst, “Towards Evolving Parametric Fuzzy Classifiers Using a Virtual Sample Generation Approach,” *16th World Congr. Int. Fuzzy Syst. Assoc. 9th Conf. Eur. Soc. Fuzzy Log. Technol.*, pp. 1111–1118, 2015.

[30] A. Narasimhamurthy and L. I. Kuncheva, “A framework for generating data to simulate changing environments,” *Proc. 25th IASTED Int. Multi-Conference Artif. Intell. Appl.*, pp. 384–389, 2007.

[31] T. S. Pham, Q. U. Nguyen, and X. H. Nguyen, “Generating Artificial Attack Data for Intrusion Detection Using Machine Learning,” *Proc. Fifth Symp. Inf. Commun. Technol.*, pp. 286–291, 2014.

[32] J. Rios-Boutin, A. Orriols-Puig, and J.-M. Garrell-Guiu, “Artificial Data Sets Based on Knowledge Generators: Analysis of Learning Algorithms Efficiency,” in *Eighth International Conference on Hybrid Intelligent Systems*, 2008, pp. 873–878.

[33] M. Takadoya, M. Notake, M. Kitahara, J. D. Achenbach, Q. C. Guo, and M. L. Peterson, “Crack-depth determination by a neural network with a synthetic training data set,” 1993.

[34] X. Zhang, Y. Fu, A. Zang, L. Sigal, and G. Agam, “Learning Classifiers from Synthetic Data Using a Multichannel Autoencoder,” *arXiv1503.03163 [cs]*, no. 1, pp. 1–11, 2015.

[35] R. Potharst and M. Van Wezel, “Generating Artificial Data with Monotonicity Constraints,” *Econ. Inst. Rep. EI 2005-06*, 2005.

[36] L.-S. Lin, D.-C. Li, and C.-W. Pan, “Improving Virtual Sample Generation for Small Sample Learning with Dependent Attributes,” *2016 5th IIAI Int. Congr. Adv. Appl. Informatics*, pp. 715–718, 2016.

[37] N. Patki, R. Wedge, and K. Veeramachaneni, “The synthetic data vault,” *Proc. - 3rd IEEE Int. Conf. Data Sci. Adv. Anal. DSAA 2016*, pp. 399–410, 2016.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[39] S. Cho, M. Jang, and S. Chang, “Virtual sample generation using a population of networks,” *Neural Process. Lett.*, vol. 5, no. 2, pp. 21–27, 1997.

[40] M. Enzweiler and D. M. Gavrila, “A mixed generative-discriminative framework for pedestrian classification,” *2008 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–8, 2008.

[41] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele, “Learning people detection models from few training samples,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1473–1480, 2011.

[42] D. Zhan and Z. Zhou, “Neighbor Line-based Locally linear Embedding.”

[43] Y. Hu *et al.*, “An Improved Algorithm for Imbalanced Data and Small Sample Size Classification,” *J. Data Anal. Inf. Process.*, vol. 3, no. 3, pp. 27–33, 2015.

[44] M. Sassano, “Virtual Examples for Text Classi cation with Support Vector Machines,” *Text*, no. 1995, 1998.

[45] S. Gaur, S. Sonkar, and P. P. Roy, “Generation of synthetic training data for handwritten Indic script recognition,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 2015–Novem, pp. 491–495, 2015.

[46] T. Varga and H. Bunke, “Generation and Use of Synthetic Training Data in Cursive Handwriting Recognition.”

[47] C. Zhang, X. Liang, and T. Matsuyama, “Generic learning-based ensemble framework for small sample size face recognition in multi-camera networks,” *Sensors (Switzerland)*, vol. 14, no. 12, pp. 23509–23538, 2014.

[48] Y. Pei and O. Zaïane, “A synthetic data generator for clustering and outlier analysis,” *Dep. Comput. Sci. Univ. Alberta …*, 2006.

[49] G. Albuquerque, T. Lowe, and M. Magnor, “Synthetic generation of

high-dimensional datasets," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2317–2324, 2011.

[50] D. Monekosso and P. Remagnino, "Synthetic Training Data Generation for Activity Monitoring and Behavior Analysis," pp. 267–275, 2009.

[51] G. Haffari, "MACHINE LEARNING APPROACHES FOR DEALING WITH LIMITED BILINGUAL TRAINING DATA IN," Simon Fraser University, 2009.

[52] R. Lanouette, J. Thibault, and J. L. Valade, "Process modeling with neural networks using small experimental datasets," vol. 23, pp. 1167–1176, 1999.

[53] D. Li, C. Wu, T. Tsai, and Y. Lina, "Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge," vol. 34, pp. 966–982, 2007.

[54] M. Wonders, Z. Ghassemlooy, and M. Alamgir Hossain, "Training with synthesised data for disaggregated event classification at the water meter," *Expert Syst. Appl.*, vol. 43, pp. 15–22, 2016.

[55] Y. Jiang, M. Li, and Z. Zhou, "Mining extremely small data sets with

application to software reuse," *Softw. - Pract. Exp.*, vol. 39, pp. 423–440, 2009.

[56] A. M. Al-bakary and S. H. Ali, "Method to Handle Data Scarcity Problem Data Construction using Genetic Programming Method to Handle Data Scarcity Problem," no. September, pp. 0–9, 2015.

[57] D. C. Li and L. S. Lin, "Generating information for small data sets with a multi-modal distribution," *Decis. Support Syst.*, vol. 66, pp. 71–81, 2014.

[58] K. Buys, J. Hauquier, C. Cagniart, T. Tuytelaars, and J. De Schutter, "Virtual data generation based on a human model for machine learning applications," *Proc. Int. Digit. Hum. Model. Conf.*, pp. 1–9, 2013.

[59] C. Wang, H. Zhao, F. Davoine, and H. Zha, "A system of automated training sample generation for visual-based car detection," *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 4169–4176, 2012.

[60] J. T. Tomás, N. Spolaôr, E. A. Cherman, and M. C. Monard, "A framework to generate synthetic multi-label datasets," *Electron. Notes Theor. Comput. Sci.*, vol. 302, pp. 155–176, 2014.

[61] J. Read, B. Pfahringer, and G. Holmes, "Generating Synthetic Multi-label Data Streams," *Ecml-Pkdd*, pp. 5–15, 2009.

[62] G. Krings, “Why synthetic data is about to become a major competitive advantage,” *Real Impact Analytics*, 2016. [Online]. Available: https://realimpactanalytics.com/en/news/why-synthetic-data-is-about-to-become-a-major-competitive-advantage. [Accessed: 01-May-2017].

[63] J. Drechsler and J. P. Reiter, “An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets,” *Comput. Stat. Data Anal.*, vol. 55, no. 12, pp. 3232–3243, 2011.

[64] J. H. Lee, I. Y. Kim, and C. O. Keefe, “On Regression-Tree-Based Synthetic Data Methods for Business Data,” no. 1, pp. 107–135, 2013.

[65] J. Manurung, H. Mawengkang, and E. Zamzami, “Optimizing Support Vector Machine Parameters with Genetic Algorithm for Credit Risk Assessment,” *J. Phys. Conf. Ser.*, vol. 930, no. 1, 2017.

[66] P. Liao, X. Zhang, K. Li, Y. Fu, M. Wang, and S. Wang, “Parameter Optimization for Support Vector Machine Based on Nested Genetic Algorithms,” *J. Autom. Control Eng.*, vol. 3, no. 6, pp. 78–83, 2015.

[67] X. Z. Li and J. M. Kong, “Application of GA-SVM method with parameter optimization for landslide development prediction,” *Nat. Hazards Earth Syst. Sci.*, vol. 14, no. 3, pp. 525–533, 2014.

[68] D. C. Li and I. H. Wen, “A genetic algorithm-based virtual sample generation technique to improve small data set learning,” *Neurocomputing*, vol. 143, pp. 222–230, 2014.

[69] J. F. Puget, “Wrong Labels In Supervised Learning,” *IBM developer Works*, 2016. [Online]. Available: https://www.ibm.com/developerworks/community/blogs/jfp/entry/Some_Limit_To_Supervised_Learning?lang=en. [Accessed: 01-May-2017].

[70] H. Reese, “Is ‘data labeling’ the new blue-collar job of the AI era?,” *TechRepublic*, 2016. [Online]. Available: http://www.techrepublic.com/article/is-data-labeling-the-new-blue-collar-job-of-the-ai-era/. [Accessed: 01-May-2017].

[71] I. Triguero, S. García, and F. Herrera, “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study,” *Knowl. Inf. Syst.*, pp. 1–40, 2013.

[72] F. Lin and W. W. Cohen, “Semi-supervised classification of network data using very few labels,” *Proc. - 2010 Int. Conf. Adv. Soc. Netw. Anal. Mining, ASONAM 2010*, pp. 192–199, 2010.

[73] J. Shin, C. Ré, and M. Cafarella, “Mindtagger: A demonstration of data labeling in knowledge base construction,” *Proc. VLDB Endow.*, vol. 8,

no. 12, pp. 1920–1923, 2015.

[74] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari, “Learning with Noisy Labels,” *Adv. Neural Inf. Process. Syst.*, pp. 1196–1204, 2013.

[75] “What Are Economic Models?” [Online]. Available: https://mostlyeconomics.wordpress.com/2011/06/16/what-are-economic-models/. [Accessed: 01-May-2017].

[76] D. Zenghelis, “What do economic models tell us?,” *Business Green*, 2013. [Online]. Available: http://www.lse.ac.uk/GranthamInstitute/news/why-economic-models-tell-us-so-little-about-the-future-2/. [Accessed: 20-Jul-2005].

[77] H. N. Barnum and L. Squire, “An econometric application of the theory of the farm-household,” *J. Dev. Econ.*, vol. 6, no. 1, pp. 79–102, 1979.

[78] L. Husin, “The Rubber Farm Household’s Behavior on Economic Activities to Achieve Family Food Security,” *APCBEE Procedia*, vol. 4, pp. 169–174, 2012.

[79] A. Bekele, K. Belay, B. Legesse, and T. Lemma, “Effects of crop commercial orientation on productivity of smallholder farmers in drought-prone areas of the central rift valley of Ethiopia.,” *J. Rural Dev.*,

vol. 33, no. 4, pp. 105–128, 2010.

[80] J. Pender and D. Alemu, “Determinants of Smallholder Commercialization of Food Crops: Theory and Evidence from Ethiopia,” *IFPRI Discuss. Pap.*, no. December, p. 88, 2007.

[81] S. Niragira, M. D’Haese, J. Buysse, S. Desiere, J. Ndimubandi, and L. D’Haese, “Options and impact of crop production specialization on small-scale farms in the Noth of Burundi,” *4th Int. Conf. African Assoc. Agric. Econ. Sept. 22-25, 2013*, 2013.

[82] R. J. Groothuis, “Decision-making of Rural Farm Households Growing Cashew or Shea,” Wageningen University, 2016.

[83] A. Adong, T. Muhumuza, and S. Mbowa, “Smallholder food crop commercialization in uganda: panel survey evidence from uganda,” no. 116, pp. 1–30, 2014.

[84] D. B. Rahut and I. V. Castellanos, “COMMERCIALIZATION OF AGRICULTURE IN THE HIMALAYAS,” *Inst. Dev. Econ.*, no. IDE Discussion Paper No. 347, pp. 1–43, 2012.

[85] E. O. Aderemi, B. T. Omonona, S. A. Yusuf, and O. A. Oni, “Determinants of output commercialization among crop farming

households in South Western Nigeria," *Am. J. Food Sci. Nutr. Res.*, vol. 1, no. 4, pp. 23–27, 2014.

[86] S. Asuming-Brempong, J. K. Anarfi, S. Arthur, and S. Asante, "Determinants of Commercialization of Smallholder Tomato and Pineapple Farms in Ghana," *Am. J. Exp. Agric.*, vol. 3, no. 3, pp. 606–630, 2013.

[87] O. U. Oteh and I. N. Nwachukwu, "Efect of Commercialization on Productive Capacity Among Cassava Producing Households in Ikwuano Local Government Area of Abia State, Nigeria," *Sci. Pap. Manag. Econ. Eng. Agric. Rural Dev.*, vol. 14, no. 3, pp. 213–220, 2014.

[88] C. Seyoum and T. Lemma, "Factors Determining the Degree of Commercialization of Smallholder Agriculture : The Case of Potato Growers in Kombolcha District , East."

[89] N. M. Agwu, C. I. Anyanwu, and E. I. Mendie, "Socio-Economic Determinants of Commercialization Among Smallholder Farmers in Abia State , Nigeria," *Greener J. Agric. Sci.*, vol. 2, no. 8, p. 10, 2013.

[90] K. Kikuhara, H. Kumagai, and H. Hirooka, "Development and Evaluation of a Simulation Model for Dairy Cattle Production Systems Integrated with Forage Crop Production," *Asian-Australasian J. Anim.*

Sci., vol. 22, no. 1, pp. 57–71, 2009.

[91] G. Kruseman and J. Bade, "Agrarian policies for sustainable land use: Bio-economic modelling to assess the effectiveness of policy instruments," *Agric. Syst.*, vol. 58, no. 3, pp. 465–481, 1998.

[92] G. Quaranta and R. Salvia, "A bio-economic model to simulate farmers behaviour in a Mediterranean desertification risky area : data needs and empirical evidences," pp. 159–168.

[93] J. M. Antle and S. M. Capalbo, "Agricultural & Applied Economics Association Econometric-Process Models for Integrated Assessment of Agricultural Production Systems Published by : Oxford University Press on behalf of the Agricultural & Applied Economics Association Stable URL : http://w," *Am. J. Agric. Econ.*, vol. 83, no. 2, pp. 389–401, 2001.

[94] W. Shang, "Research on Chinese Rural GDP Forecasting Using Grey Model Optimized by," in *International Conference on Industrial and Information Systems*, 2009, pp. 218–221.

[95] I. A. W. Mohamed, "Introduction to the Macroeconomic Structure of Yemen," Khartoum, 2011.

[96] A. Sixtus, Ibekwe, and A. Olufemi, "Agriculture Share of the Gross

Domestic Product and its Implications for Rural Development," 2010.

[97] K. E. Eravwoke and A. Ukavwe, "An Empirical Analysis of Cash Crop Production, Transport and the Nigerian Economy\n," *IOSR J. Humanit. Soc. Sci.*, vol. 19, no. 12, pp. 22–28, 2014.

[98] M. Zulfiqar and A. F. Chishti, "Development of Supply and Demand Functions of Pakistan ' s Wheat Crop," vol. 1, no. Summer, pp. 91–102, 2010.

[99] G. C. Nelson *et al.*, "Climate change effects on agriculture: economic responses to biophysical shocks.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 9, pp. 3274–9, 2014.

[100] P. Esteve, C. Varela-ortega, I. Blanco-gutiérrez, and T. E. Downing, "A hydro-economic model for the assessment of climate change impacts and adaptation in irrigated agriculture," *Ecol. Econ.*, vol. 120, pp. 49–58, 2015.

[101] H. Larsson, "A crop loss model and economic thresholds for the grain aphid , Sitobion avenae (F .), in winter wheat in southern Sweden," *Crop Prot.*, vol. 24, pp. 397–405, 2005.

[102] A. Choudhury and J. Jones, "CROP YIELD PREDICTION USING

TIME SERIES," vol. 15, no. 3, pp. 53–68, 2014.

[103] M. García-vila and E. Fereres, "Combining the simulation crop model AquaCrop with an economic model for the optimization of irrigation management at farm level," *Eur. J. Agron.*, vol. 36, pp. 21–31, 2012.

[104] D. H. Freedman, "Why Economic Models Are Always Wrong," *Scientific American*, 2011. [Online]. Available: https://www.scientificamerican.com/article/finance-why-economic-models-are-always-wrong/. [Accessed: 01-May-2017].

[105] A. Shah, "Machine Learning vs Statistics," *KDnuggets*, 2016. [Online]. Available: http://www.kdnuggets.com/2016/11/machine-learning-vs-statistics.html. [Accessed: 01-May-2017].

[106] "ASSUMPTIONS IN EMPIRICAL ECONOMICS," *Pieria*, 2014. [Online]. Available: http://www.pieria.co.uk/articles/assumptions_in_empirical_economics. [Accessed: 01-May-2017].

[107] T. SRIVASTAVA, "Difference between Machine Learning & Statistical Modeling," *Analystics Vidhya*, 2015. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/. [Accessed: 01-May-2017].

[108] H. Sinha, “Machine Learning VS. Statistical Modeling: The Real Difference,” *infogix*, 2015. [Online]. Available: http://www.infogix.com/blog/machine-learning-vs-statistical-modeling-the-real-difference/. [Accessed: 01-May-2017].

[109] L. Maria Badea, “Predicting Consumer Behavior with Artificial Neural Networks,” *Procedia Econ. Financ.*, vol. 15, no. 14, pp. 238–246, 2014.

[110] J. Shahrabi, J.; Mousavi, S.S.; Meydar, “Supply Chain Demand Forecasting: A Comparision of Machine Learning Techniques and Traditional Methods,” *Journal of Applied Sciences*, vol. 9, no. 3. pp. 521–527, 2009.

[111] T. Raffinot, “Investing through Economic Cycles with Ensemble Machine Learning Algorithms,” vol. 2016, no. May, pp. 1–31, 2016.

[112] S. H. Chai and J. S. Lim, “Economic Turning Point Forecasting Using Neural Network with Weighted Fuzzy Membership Functions,” in *20th Interntional Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 2007, pp. 145–154.

[113] H. Ahir and P. Loungani, “‘There will be growth in the spring’: How well do economists predict turning points?,” *CEPR’s Policy Portal*, 2014. [Online]. Available: http://voxeu.org/article/predicting-economic-

turning-points. [Accessed: 01-May-2017].

[114] L. Mora-Lopez, M. Piliougine, J. E. Carretero, and M. Sidrach-De-Cardona, “Integration of statistical and machine learning models for short-term forecasting of the atmospheric clearness index,” in *International Congress on Environmental Modelling and Software*, 2010, vol. 3.

[115] L. Xiang, Y. Zhu, and G. Tang, “A Hybrid Support Vector Regression for Time Series Forecasting,” in *World Congress on Software Engineering*, 2009, vol. 4, pp. 161–165.

[116] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University Michigan of Press, 1975.

[117] M. Obitko, “Introduction to Genetic Algorithms,” 1998. [Online]. Available: http://www.obitko.com/tutorials/genetic-algorithms/selection.php. [Accessed: 01-May-2017].

[118] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston: Addison-Wesley Longman Publishing Co, 1989.

[119] C. Stephens and H. Waelbroeck, “Schemata Evolution and Building

Blocks," Mexico, 1997.

[120] M. D. Vose, *The Simple Genetic Algorithm: Foundations and Theory*. Cambridge: MIT Press, 1998.

[121] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell, "Finite populations induce metastability in evolutionary search," *Phys. Lett. A*, vol. 229, no. 3, pp. 144–150, 1997.

[122] "Gradualism and Punctuated Equilibrium," *NECSI*. [Online]. Available: http://necsi.edu/projects/evolution/evolution/grad+punct/evolution_grad+punct.html. [Accessed: 01-May-2017].

[123] "More on punctuated equilibrium," *Understanding Evolution*. [Online]. Available: http://evolution.berkeley.edu/evolibrary/article/side_0_0/punctuated_01. [Accessed: 01-May-2017].

[124] V. N. Vapnik, *Statistical Learning Theory*. Wiley, 1998.

[125] "3. The Law of Large Number." [Online]. Available: http://www.math.uah.edu/stat/sample/LLN.html. [Accessed: 01-May-2017].

[126] "Structural Risk Minimization." [Online]. Available:

http://www.svms.org/srm/. [Accessed: 01-Jul-2017].

[127] A. Blum, “Machine Learning Theory.”

[128] S. Haylin, *Neural Networks A Comprehensive Foundation*, 2nd ed. New Jersey: Prentice-Hall Inc, 1999.

[129] F. I. García, “Introduction to Support Vector Machines,” *OpenCV*. [Online]. Available: https://docs.opencv.org/2.4/doc/tutorials/ml/table_of_content_ml/table_of_content_ml.html. [Accessed: 01-Jul-2017].

[130] T. Fletcher, “Support Vector Machines Explained,” *Online]*. *http//sutikno. blog. undip. ac. id/files/2011/11/SVM-Explained. pdf.[Accessed 06 06 2013]*, pp. 1–19, 2009.

[131] N. Kantanantha, “Crop Decision Planning Under Yield and Price Uncertainties,” Georgia Institute of Technology, 2007.

[132] M. Q. Dao, “Population and economic growth in developing countries.,” *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 2, no. 1, pp. 6–17, 2012.

[133] E. Kantar and M. Keskin, “The relationships between electricity consumption and GDP in Asian countries, using hierarchical structure methods,” *Phys. A Stat. Mech. its Appl.*, vol. 392, no. 22, pp. 5678–5684,

2013.

[134] M. A. Beyzatlar, M. Karacal, and H. Yetkiner, “Granger-causality between transportation and GDP : A panel data approach,” *Transp. Res. Part A*, vol. 63, pp. 43–55, 2014.

[135] L. Zeng, “Evaluating GDP Forecasting Models for Korea,” 2011.

[136] K. Barhoumi *et al.*, “Short Term Forecasting of GDP using Large Monthly Datasets: A Pseudo Real Time Forecast Evaluation Exercise,” 2008.

[137] M. Marcellino, “A comparison of time series models for forecasting GDP growth and inflation,” Milan, 2007.

[138] “Agriculture’s Contribution to Rural Development,” *International Conference on Non-Trade Concerns in Agriculture*, 2000. [Online]. Available: https://www.iatp.org/files/Agricultures_Contribution_to_Rural_Development.htm. [Accessed: 01-May-2017].

[139] M. Duncan, “Farm Real Estate Values - Some Important Determinants,” Kansas, 1977.

[140] A. B. Kamara, “The impact of market access on input use and

agricultural productivity: evidence from Machakos district, Kenya," *Agrekon*, vol. 43, no. 2, 2004.

[141] A. Dorward *et al.*, "Protocol: Agricultural Input Subsidies for improving Productivity, Farm Income, Consumer Welfare and Wider Growth in Low- and Middle-Income Countries: A Systematic Review," no. August 2013, pp. 1–45, 2014.

[142] G. Guo, Q. Wen, and J. Zhu, "The Impact of Aging Agricultural Labor Population on Farmland Output: From the Perspective of Farmer Preferences," *Math. Probl. Eng.*, vol. 2015, 2015.

[143] L. Musemwa, A. Mushanje, V. Muchenje, F. Aghdasi, and L. Zhou, "Factors affecting efficiency of field crop production among resettled farmers in Zimbabwe," *Int. Conf. African Assoc. Agric. Econ.*, no. September 22-25, 2013, 2013.

[144] S. W. Foli, "ITS IMPACT ON HOUSEHOLD FARM INCOME : THE CASE OF," HARAMAYA UNIVERSITY, 2013.

[145] U. N. Opara, "Personal and Socio-Economic Determinants of Agricultural Information Use by Farmers in the Agricultural Development Programme (ADP) Zones of Imo State, Nigeria," 2010.

[146] J. C. Anyanwu, “Marital Status, Household Size and Poverty in Nigeria: Evidence from the 2009/2010 Survey Data,” *African Dev. Rev.*, vol. 26, no. 1, pp. 118–137, 2014.

[147] I. K. Okpachu, A.S., Okpachu, O.G., Obijesi, “The impact of education on agricultural productivity of small scale rural female maize farmers in Potiskum Local Government , Yobe State : A panacea for rural economic development in Nigeria .,” *Int. J. Res. Agric. Food Sci.*, vol. 2, no. 4, pp. 26–33, 2014.

[148] Z. Xiong and Y. Niu, “Factors Affecting the Income of Farmers,” *Asian Agric. Res.*, vol. 2, no. 5, pp. 18–26, 2010.

[149] M. S. Bhatt and S. A. Bhat, “TECHNICAL EFFICIENCY AND FARM SIZE PRODUCTIVITY— MICRO LEVEL EVIDENCE FROM JAMMU & KASHMIR,” *Int. J. Food Agric. Econ. ISSN*, vol. 2, no. 4, pp. 2147–8988, 2014.

[150] A. W. Masunga, “ASSESSMENT OF SOCIO-ECONOMIC AND INSTITUTIONAL FACTORS INFLUENCING TOMATO PRODUCTIVITY AMONGST SMALLHOLDER FARMERS: A CASE STUDY OF MUSOMA MUNICIPALITY, TANZANIA,” Sokoine University of Agriculture, 2014.

[151] D. Birkhaeuser and R. E. Evenson, “The Economic Impact of Agricultural Extension : A Review,” *Econ. Dev. Cult. Change*, vol. 39, no. 3, pp. 607–650, 1991.

[152] M. A. Baloch and G. B. Thapa, “The effect of agricultural extension services: Date farmers’ case in Balochistan, Pakistan,” *J. Saudi Soc. Agric. Sci.*, 2016.

[153] W. C. Found, “Towards a General Theory Relating Distance between Farm and Home to Agricultural Production,” *Geogr. Anal.*, vol. 2, no. 2, pp. 165–176, 1970.

[154] E. O. Obiero, “Social Economic Factors Affecting Farm Yield in Siaya District, Siaya County, Kenya,” p. 59, 2013.

[155] M. T. Parvin and M. Akteruzzaman, “Factors Affecting Farm and Non-Farm Income of Haor Inhabitants of Bangladesh,” *Factors Affect. Farm Non-Farm Income Haor Inhabitants Bangladesh*, vol. 23, pp. 143–150, 2012.

[156] C. Gerstter, T. Kaphengst, D. Knoblauch, and K. Timeus, “An assessment of the effects of land ownership and land grab on development, with a particular focus on small holdings and rural areas,” p. 32, 2011.

[157] S. Tenaw, K. M. Z. Islam, and T. Parviainen, “Effects of land tenure and property rights on agricultural productivity in Ethiopia, Namibia and Bangladesh,” p. 33, 2009.

[158] I. Kan, A. Kimhi, and Z. Lerman, “Farm output , non-farm income and commercialization in rural,” *J. Agric. Dev. Econ.*, vol. 3, no. 2, pp. 276–286, 2006.

[159] A. G. Doti, “Causes and Effects of Land Size Variation on Smallholder’s Farm-Income: The Case of Kombolcha District of East Hararghe, Oromia, Ethiopia,” *OALib*, vol. 4, no. 1, pp. 1–17, 2017.

[160] P. Amaza and G. Ayoola, “Enhancing the Access of Rural Households to Output Markets for Increased Farm Incomes,” 2006.

[161] P. O. Agbola, A. Thomas, and B. D.A, “Determinants of farmers’ access to output markets and the effects on income; A case study of Ikenne local government area, Nigeria,” *J. Life Phys. Sci.*, vol. 3, no. 2, pp. 33–39, 2010.

[162] A. Barry, “Weather causing problems for farmers,” *thejournal.ie*. [Online]. Available: http://www.thejournal.ie/weather-farmers-problems-ireland-strawberries-510509-Jul2012/. [Accessed: 01-May-2017].

[163] “Measures of Central Tendency: Mean, Median, and Mode,” *Owlcation*, 2014. [Online]. Available: https://owlcation.com/stem/Measures-of-Central-Tendency-Mean-Median-and-Mode. [Accessed: 01-May-2017].

[164] “Measures of Central Tendency,” *Laerd Statistics*. [Online]. Available: https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php. [Accessed: 01-May-2017].

[165] Y. Duan, “Household Formation and Composition in China’s Ethnic. Minority Populations: A Case Study of Six Yunnan Villages ,” p. 12.

[166] J. Ledgerwood, “Rural Development in Cambodia: The View from the Village,” *Cambodia and the International Community: The Quest for Peace, Development, and Democracy*. pp. 127–147, 1998.

[167] T. Pasanen, “Multidimensional Poverty in Laos: Analysis on Household and Village Levels,” *J. Int. Dev.*, 2016.

[168] “Global Health Observatory (GHO) data,” *World Health Organization*. [Online]. Available: http://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends/en/. [Accessed: 01-May-2017].

[169] “Malaysia,” *Education Policy and Data Center*, 2014. [Online].

Available: https://www.epdc.org/country/Malaysia. [Accessed: 01-May-2017].

[170] "Education in Nigeria," *Maps of World*. [Online]. Available: https://www.mapsofworld.com/nigeria/education/. [Accessed: 01-May-2017].

[171] "Education," *UNICEF Nigeria*. [Online]. Available: http://www.unicef.org/nigeria/children_1937.html. [Accessed: 01-May-2017].

[172] S. Pandey, H. Bhandari, and B. Hardy, *Economic costs of drought and rice farmers' coping mechanisms: a cross-country comparative analysis*, vol. 53. 2007.

[173] S. K. Lowder, J. Skoet, and T. Raney, "The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide," *World Dev.*, vol. 87, pp. 16–29, 2016.

[174] T. Takane, "Labor use in smallholder agriculture in Malawi: six village case studies," *Afr. Study Monogr.*, vol. 29, no. 4, pp. 183–200, 2008.

[175] T. S. Jayne *et al.*, "Smallholder income and land distribution in Africa: Implications for poverty reduction strategies," *Food Policy*, vol. 28, no.

3, pp. 253–275, 2003.

[176] D. I. J. Ha Dinh Tuan, Nguyen Ngoc Hue, Bhuwon R. Sthapit, “On-farm management of agricultural biodiversity in Vietnam,” in *International Plant Gentic Resources Institute*, 2001.

[177] R. V. Gerpacio, J. D. Labios, R. V. Labios, and E. I. Diangkinay, *Maize in the Philippines: Production Systems, Constraints, and Research Priorities*. 2004.

[178] S. Wanmali, “Rural infrastructure, the settlment system, and development of the regional economy in Southern India,” *Ifpri*, no. RR 91. 1992.

[179] “Land Requirement for Agricultural Enterprises: Guidelines,” in *Agricultural Guidelines*, .

[180] “Agricultural Resource Management Survey, Phase 2, 1996-2010.,” 2015.

[181] R. Naylor, “THE RURAL LABOR MARKET IN INDONESIA.”

[182] M. Duffy, “Estimated Costs of Crop Production in Iowa - 2014,” *Iowa State Univ. Ext. Outreach*, vol. A1-20, no. January, pp. 1–13, 2014.

[183] “Village households’ average assets at Rs 10 lakh,” *The Economic*

Times, 2014. [Online]. Available: http://economictimes.indiatimes.com/news/economy/indicators/village-households-average-assets-at-rs-10-lakh/articleshow/45578614.cms. [Accessed: 01-May-2017].

[184] M. J. Campbell, Ed., *New Technology and Rural Development*. The Association of Development Research and Training Institutions of Asia and the Pacific.

[185] "Kernel." [Online]. Available: http://blog.pluskid.org/?p=685. [Accessed: 01-Sep-2017].

[186] Y. Ming-hui and W. Ru-chuan, "DDoS detection based on wavelet kernel support vector machine," *J. China Univ. Posts Telecommun.*, vol. 15, no. 3, pp. 59–63, 2008.

[187] L. Tang, L. Tang, and H. Sheng, "Forecasting volatility based on wavelet support vector machine," *Expert Syst. Appl.*, vol. 36, pp. 2901–2909, 2009.

[188] L. Zhang, W. Zhou, and L. Jiao, "Wavelet Support Vector Machine," *Syst. MAN, Cybern. B Cybern.*, vol. 34, no. 1, pp. 34–39, 2004.

[189] E. W. Weisstein, "Sinc Function," *From MathWorld--A Wolfram Web*

Resource. [Online]. Available: http://mathworld.wolfram.com/SincFunction.html. [Accessed: 01-Sep-2017].

[190] J. George and K. Rajeev, "Hybrid Wavelet Support Vector Regression," in *Cybernetic Intelligent Systems*, 2008.

[191] R. Ghiasi, P. Torkzadeh, and M. Noori, "A machine-learning approach for structural damage detection using least square support vector machine based on a new combinational kernel function," *Struct. Heal. Monit.*, vol. 15, no. 3, pp. 302–316, 2016.

[192] B. A. Sonkamble and D. D. Doye, "A Novel Linear-Polynomial Kernel to Construct Support Vector Machines for Speech Recognition Pune Institute of Computer Technology , Pune , Maharashtra State , India Department of E and TC , SGGSIE and T , Nanded , Maharashtra State , India," *J. Comput. Sci.*, vol. 7, no. 7, pp. 991–996, 2011.

[193] P. R. Fisk, "The Graduation of Income Distributions," *Econometrica*, vol. 29, no. 2, pp. 171–185, 2016.

[194] C. Kleiber and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*. New Jersey: Wiley-Interscience, 2003.

[195] C. Mccarter, “10-601B Recitation 11,” pp. 1–7, 2015.

[196] T. Van Gestel, B. Baesens, J. A. K. Suykens, D. Van den Poel, D. E. Baestaens, and M. Willekens, “Bayesian kernel based classification for financial distress detection,” *Eur. J. Oper. Res.*, vol. 172, no. 3, pp. 979–1003, 2006.

[197] “Advantages and problems of contract farming.” [Online]. Available: http://www.fao.org/docrep/004/y0937e/y0937e03.htm.

[198] “The World Bank Atlas method - detailed methodology,” *The World Bank*. [Online]. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/378832-the-world-bank-atlas-method-detailed-methodology. [Accessed: 01-May-2017].

[199] “What is the SDR deflator?,” *The World Bank*. [Online]. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/378829-what-is-the-sdr-deflator.

[200] “World Bank Country and Lending Groups.” [Online]. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519.

[201] H. Wu, Y. Ning, P. Chakraborty, J. Vreeken, N. Tatti, and N.

Ramakrishnan, “Generating Realistic Synthetic Population Datasets,” *Comput. Res. Respository*, pp. 1–16, 2016.

Appendix A: Report on Using Synthetic Data Generated from GA for Support Vector Regression Training to Predict Cost of Production of Underutilised Crop

Regression is one of the popular methods used in machine learning and statistics. In this experiment, Support Vector Regression (SVR) is used to predict the cost of Bambara nuts. To do this, data of an equivalent crop was needed as the training data and data of Bambara nuts were used as the test data. A few characteristics are set to match the Bambara nuts in the search of the equivalent crop. Table 1 shows the comparison between Bambara nut and peanut.

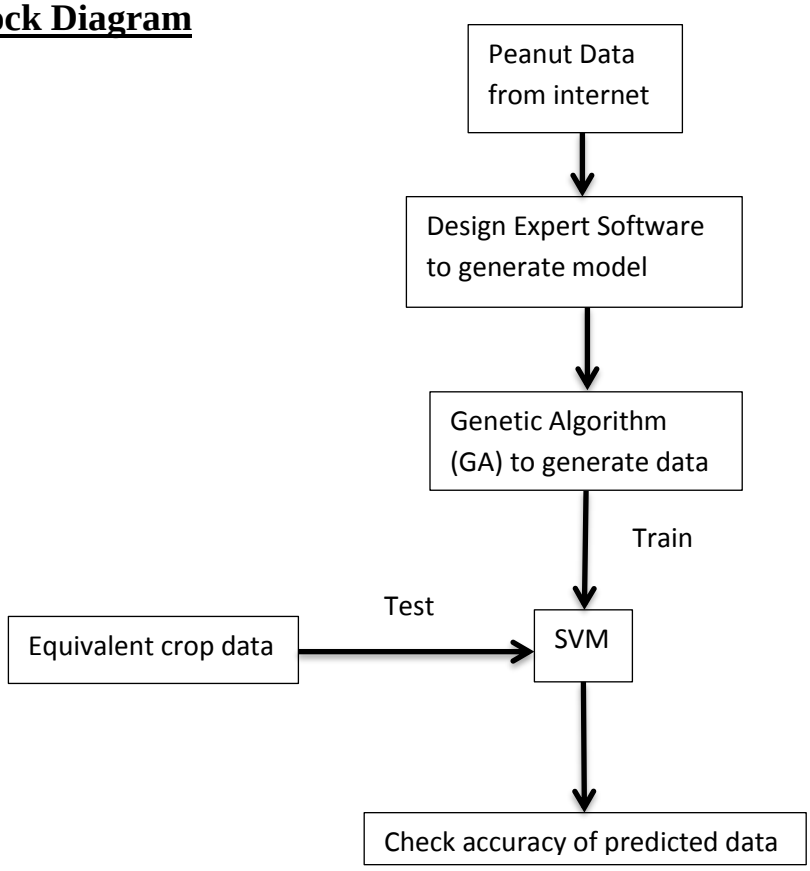
Bambara nut	**Characteristics**	**Peanut**
5 – 6.5	**pH of Soil**	5.9 – 7
90 – 170 days	**Growth Cycle**	90 – 150 days
Light sandy soil	**Type of Soil**	Light sandy soil
700mm – 1400mm	**Average rainfall**	500mm – 1500mm
Tropical wet and dry	**Climate**	Warm and moist condition

Table 1

Methodology

1) Peanut data was collected from the internet.
2) A model is generated using Design Expert Software from the dataset of peanut data.
3) The generated model is used as cost function in Genetic Algorithm to create thousands of data points, constraints were set to reflect the real data.
4) The generated data is used as the training data to train SVR.
5) Another equivalent crop data is obtained from the internet as the test data.
6) The MSE and Squared Correlation Coefficient is recorded.

Block Diagram



Data Collection

Peanut plant is chosen as the equivalent crop, the peanut farm cost was searched across the internet and collated according to the specified parameters. 52 sets of data were found from Texas peanut farm. Table 2 shows the list of parameters required for the regression. The irrigation equipment and tractor/machinery parameters are the cost of license/rental required to run these equipments. For new farmers these two parameters can be input as the cost of purchasing the equipment/tractor/machinery.

Land Rental	Seed Cost	Labor	Fertilizer
Irrigation equipment and installation	Irrigation Fuel	Tractor and Machinery	Repair and Maintenance of Machinery/Equipment
Herbicide	Insecticide and Fungicide	Fuel for Tractor and Machinery	Drying & Cleaning

Table 2

Problems with the Equivalent Plant Data

The data was not sufficient to train SVM. To solve this problem, Genetic Algorithm (GA) is used to generate hundreds or thousands of artificial data from the raw input data.

Curve Fitting and Genetic Algorithm

52 points of data are used to generate a linear model using Design Expert Software. The model is developed as follows:

Total cost = -219.05054 + 0.96584*(land rental) + 0.78148*(seed cost) + 1.15620* (labor) + 0.83327*(fertilizer) + 4.69986*(Herbicide) - 0.41386*(Insecticide) + 1.22383*(Irrigation Installation) + 1.10188*(Irrigation Energy) + 0.016652*(Tractor and Machinery) + 1.17803*(Repair and Maintenance) + 1.71190*(Fuel) + 2.13437*(Dry& Clean)

This model is then used as cost function of GA to generate artificial data.

Applying Regression using GA data

20% of the peanut data is used as test data to validate the model. 4 different kernels were used. The results were recorded in Table 3 and 4.

Kernel Function	Mean Squared Error (MSE)	Squared correlation coefficient (R^2)
Linear	190165	0.997441
Polynomial	815597	0.971528
Radial Basis Function(RBF)	4.84084e+06	5.93995e-16
Sigmoid	5.94498e+06	-2.67298e-15

Table 3

Real Cost	Linear	Polynomial	RBF	Sigmoid
4440.50	4942.37	4815.31	5586.53	5850.56
4878.39	5250.87	5227.04	5586.53	5850.56
4479.36	4944.40	4803.72	5586.53	5850.56
3452.68	3920.04	4263.48	5586.53	5850.56
4377.84	4762.57	4938.65	5586.53	5850.56
4702.39	5133.10	5288.52	5586.53	5850.56
2717.97	3049.31	3764.12	5586.53	5850.56
2224.70	2697.29	3680.58	5586.53	5850.56
2530.16	2975.79	3779.27	5586.53	5850.56
2472.09	2931.55	3775.53	5586.53	5850.56

Table 4

Discussion

From the table above, larger MSE and smaller R^2 is observed for RBF and sigmoid kernel as compared to Polynomial and Linear kernels. The small R^2 shows that the predicted cost varies from the real costs. The predicted cost are the same cost for each case. This is because the regression line using RBF or Sigmoid function might over fitted the data and ignored the trend of the data which ends up predicting the same for all input cases. The polynomial function has a smaller error and larger R^2 compare to RBF and Sigmoid but the average percentage difference between predicted cost and real cost is approximately 27.7%. This difference may be caused by the polynomial line does not fit well and it might have some slight variation in predicting the cost. For the linear function, it has the smallest error and the R^2 is very close to 1. This shows that the predicted costs do not vary much from the real cost. Although polynomial's R^2 difference by 0.02 from linear, a small change in R^2 will change the accuracy

of the predicted cost and the average percentage difference is around 13%. This is because the equation generated from Design Expert Software is in linear form. Hence, during GA it generated the data linearly and it suits well for the regression in linear setting.

Equivalent crop vs. non-equivalent crop

In order to further validate the methodology, data of an equivalent crop and a non-equivalent crop is used as test data. Pinto bean was chosen as the equivalent crop due to the similarity in characteristics as peanut shown in Table 5. Cotton was picked to represent as non-equivalent crop.

Characteristics	Peanuts	Pinto Beans	Cotton
pH of Soil	5.9 – 7	6.0 – 7.0	5.8 – 8.0
Growth Cycle	90 – 150 days	90 – 150 days	180 – 195 days
Type of Soil	Light sandy soil	Light sandy loam	Sandy Loam
Average rainfall	500mm – 1500mm	600mm – 1400mm	700mm – 1300 mm
Climate	Warm and moist condition	Warm condition	Subtropical and tropical Regions

Table 5

Although cotton has some similar traits as the other two crops, the pH and growth cycle of cotton are different from the two. This will influence the parameter of farm cost, thus cotton is non-equivalent crop for peanut. Data of both crops used as test data while the training data is still the data generated from GA using peanut cost model. Linear kernel was used. Table 6 shows the result of pinto beans and cotton.

Pinto Beans		Cotton	
Mean Squared Error : 233590		Mean Squared Error : 1.96645e+06	
Squared correlation coefficient: 0.992265		Squared correlation coefficient: 0.892365	
Real Cost	Predicted Cost	Real Cost	Predicted Cost

3081.60	3554.42	4052.99	5483.42
2568.00	3001.26	4156.86	5596.23
2465.02	3026.55	4024.53	5596.23
2636.48	3102.33	3751.12	5384.76
2824.80	3234.50	3447.24	4558.48
1600.72	2053.88	3421.39	4802.65
2396.80	2900.62	3541.06	5021.54
2054.40	2596.85	3374.65	4777.96
1712.00	2226.82	3617.41	4874.30
1455.20	1909.02	3362.88	4597.39

Table 6

From the table it can be seen that the MSE of pinto beans is smaller than cotton. The R^2 of pinto beans is much closer to 1 than cotton. The average percentage difference between pinto beans and cotton is approximately 22% and 38% respectively. Comparing the results obtain using peanuts data, pinto beans average percentage difference is increased by 9% while cotton is increased by 25%. This increase of the average percentage difference is because pinto beans usually do not dry after harvesting while peanuts is dried/clean after harvesting. This means the beans were freshly harvested and send to the factory/market to sell them while peanuts have to dry and clean before sending them to the factory. Therefore, SVR predicted the cost of the pinto beans with the drying/cleaning process included. For the large average percentage difference case of cotton, the regression model of the training data is more suitable for the peanut/equivalent crop.

Conclusion

As a conclusion, this methodology is able to predict farm cost close to real cost. GA is used to generate artificial data from equivalent crop model in order to train SVM when real training data is not sufficient or totally unavailable to make accurate prediction. It is recommended to have larger sets of training data and the parameters to be revised so that more data are able to fit in. With more input data, the predicted cost will be more accurate and precise. Using peanuts to predict pinto beans gave a 22% difference from the real pinto beans data.