Classification Algorithm for Avian Influenza Surveillance Model of Poultry Farms in Thailand

Sirapote Santatiwut [1], *Anongnart Srivihok* [2]
Department of Computer Science, Faculty of Science, Kasetsart University, Thailand
email: g4964166@ku.ac.th[1], fsciang@ku.ac.th[2]

Abstract

The pandemic of Avian Influenza (AI) in Thai poultry farms have threatened the growth of agriculture industry in Thailand. Since more than thousand poultries were dead and eradicated due to AI infections. The purpose of this study is to find the algorithm which can be used to build a classification model that predicts the site of poultry farms which are positive (infected) AI. Two classification algorithms: C4.5 and ID3 were tested on the prediction accuracy, precision and recall. The factors using for model generation including farm region, site characteristics, temperature, month, season, number of poultry, humidity, numbers daylight hours, numbers of sick poultries, numbers of dead and numbers of destroyed poultries.

Results showed that Decision Tree (C4.5) should be applied for Avian Influenza Surveillance (AIS) model. This proposed model is beneficial for predicting the positive site for AI epidemic. Thus the officials or farmers can use this model for AIS and protection of AI pandemic in the future.

Keyword: Classification, Decision Tree, Avian Influenza Surveillance

1. Introduction

At the present, the crisis of the bird Flu is critical as the important outbreak in the global level that leads to several disasters such as economic, health, life quality and agriculture.

There are the serious pandemic of the Bird Flu, especially in Thailand. Journalists report Avian Influenza infections in poultry farms and some examples on human being in daily news.

The Avian Influenza (Bird Flu) is the name of virus which classified as the flu type H5N1, which is infected in poultry farm. It also can transfer to people through touching with sick poultry or infected animal corpses.

In Thailand, the serious of Bird Flu is considerable to be severe disease due to Thailand's economic is based on agriculture industry. In the past Bird Flu spreads rapidly and there were many infections in various zones. Researchers attempt to study on Bird Flu pandemic. This might provide information for efficiently protect and control the epidemic of disease, warning and watch out the danger of the Avian Influenza

In this paper, the classification model of Avian Influenza was proposed. The model was created by using a Decision Tree algorithm and the measurement of model efficiency by using accuracy rate, precision and recall.

2. Related Theory

2.1 Decision Tree

Decision Tree is an algorithm which uses for the Data Mining and Classification by creating a prediction model showing the relationship between parameters and results. Decision Tree's structure consists of Leave Nodes expressed classification's result and branch of Decision Tree that shown the linking of parameters classified for results of Leave Node.

There are many types of Decision Trees. In this study, two decision tree algorithms: ID3 and C4.5 were applied since they are most widely used in classification models.

ID3 [1] is an algorithm used for creating of a Decision Tree. This algorithm uses the concept of Information Entropy as:

$$I_E(i) = -\sum_{j=1}^{m} f(i,j)\log_2 f(i,j)$$

Similarly, C4.5 [8] is an algorithm used for creating of Decision Tree which was developed from ID3. The successor, C4.5 is using to predict classes in both numeric and non-numeric. However, C4.5 has more advantages as follows:

1. It can be used with both continuous and discrete parameters.
2. It complied with uncompleted parameters.
3. It is pruned after Decision Tree is generated.

2.2 Related Studies

Abbas and coworkers proposed the model of influenza in order to analyze time and patients who potentially have risk to be infected by flu in the spreading period. They tested hypothesis whether flu vaccine was capable to control flu well or not. Results showed that the vaccine performance was depending on its intense and efficiency [2].

Later, Abbas and coworkers used Bay Net to create a model to analyze how accurate the population was infected by the spread of flu [3]. Cooper and et al. [4] studied Causal Bayesian network and proposed a model: Spatial-temporal. This model was used to watch out and measure air-spreading epidemic disease which can infect without contacting infected sources.

Li and his researchers [5] built a Transmission Model which used for explaining the spread of Bird Flu and predicting the future outbreak in mainland China. The outcome of this research is the model used for predicting infected areas in China.

Upadhyay and coworkers [6] focused on the spread of Avian Influenza in the terms of time and area in order to know how it was spread. This causes the model used for the efficient surveillance.

Sebastiani and group's research [7] presented the effects of flu by Dynamic Bayesian Network. This was used as a tool for the disease warning and the study of the express of patients who were sick in childhood with Severe Acute Respiratory Syndrome (SARS). These patients were treated in hospital's baby-care program for making the index of outbreak rates and death rates.

Hak and coworkers studied the model creation in order to monitor the side effects in Health – Economic area where flu spreads in the future. By using Decision Tree and basic data of the flu outbreak in the Netherlands, the model was proposed [9].

From the above studies, it is obviously seen that researchers use Bayesian Network to present the outbreak of flu and other diseases. However, the study of Avian Influenza in Thailand, which is one type to Flu, is never conducted by using data mining techniques. It is the purpose of this study to create a classification model of poultry farms to be infected by Avian Influenza.

3. Experiments

Data sets were obtained from the report of infected death of poultry since July 2004-2007 (the Bureau of Livestock) and weather reports of the same period from the Meteorological Department. There are totally about 8,501 records which classified as death proof from (1) the Bird Flu 1,647 records and (2) another reason, but not Bird Flu 6,854 records. Experimental Data were classified into 2 classes: (1) positive class which means dead poultry in farms infected by Bird flu and, (2) negative class which means dead poultry in farms not infected by Bird Flu.

There were 12 parameters in the model generating including (1) zone (2) month (3) season (4) Temperature (5) Humidity (6) Sun Light Day Hours (7) Number of poultry (8) number of sick poultry (9) number of death poultry (10) number of poultry was Destroyed (11) Type of poultry and (12) region as shown in Table 1.

To measure model's efficiency, RMSE (Root Mean Square Error), Recall, Precision and F-measure of the tested models were used for evaluations

Table 1 shows attributes of data in the model creation.

Name of variables	Attribute Type
Region	Nominal
Zone	Nominal
Month	Nominal
Season	Nominal
Temperature	Numeric
Humidity	Numeric
Sun light Day Hour	Numeric
Number Of Poultry	Numeric
Number of Death Poultry	Numeric
Number of Sick Poultry	Numeric
Number of poultry was Destroyed	Numeric
Type Poultry	Category

Three steps in the model creation are as follows:

Step 1. Data Preprocessing.
Data were collected from reports of infected death of poultry between July 2004-2007 (the Bureau of Livestock) and the weather reports of the same periods from the Meteorological Department of Thailand. Uncompleted data and missing data were deleted. Then data files were formatted to CSV form for data mining software, WEKA version 3.4.13.

Step 2 Classification Models.
Create two classification models by using C4.5 and ID3 algorithms. Data set was obtained from step 1.

Step 3. **Evaluation.**
The measurements of model performances were conducted by using 10-fold cross validation. Then, precision, recall, F-Measure, accuracy rate and RMSE were used to test model efficiency. The comparisons of efficiency between two models (C4.5. and ID3) were conducted.

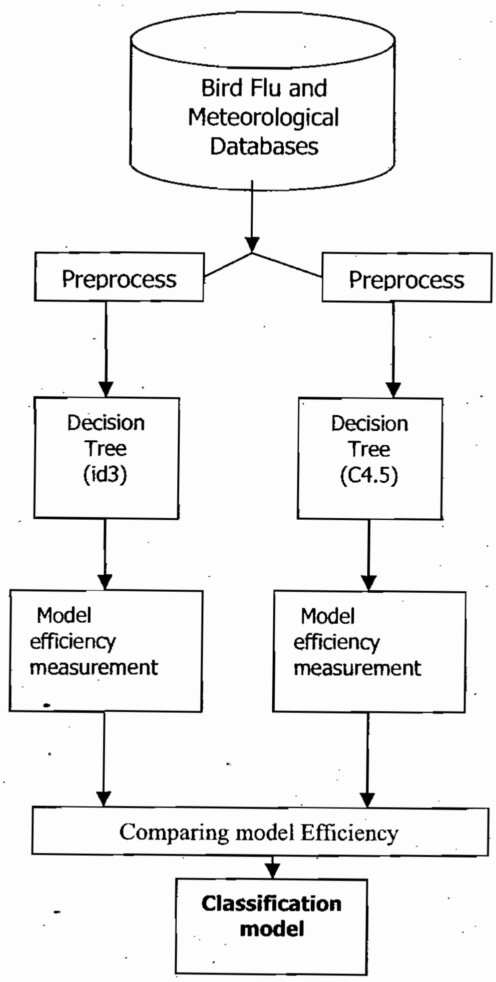

Figure 1. Framework of the study.

4. Results

Results in table 2 showed that the prediction power of DT model (C4.5) is good with high correctly classified (91.11%) and low RMSE (0.2629). From Table 3, results of C4.5 predictions for Positive class, precision is 0.754, recall is 0.688 and F-measure is 0.72. Whereas, Negative class, precision is 0.939, recall is 0.955 and F-measure is 0.94 which is very good. However, Table 3 shows the prediction power of DT model (ID3) has lower accuracy rate (83.488%) than C4.5. Further ID3 provides the higher RMSE (0.3185) and lower precision and recall value than C4.5. Thus C4.5 algorithm has higher performances than ID3. Then, C4.5 was selected to generate the decision tree for AI model

Table 2 shows the comparisons of efficiency of C4.5 and ID3

Algorithm	Correctly Classified	Incorrectly classified	RMSE
C4.5	91.11	8.89	0.2629
ID3	83.4887	11.6045	0.3185

Table 3 shows the comparisons of Precision, Recall, F-measure of model created by Decision Tree in two classes

Algorithm	Precision	Recall	F-measure	Class
C4.5	0.754	0.688	0.72	Pos
	0.939	0.955	0.947	Neg
ID3	0.632	0.49	0.929	Pos
	0.911	0.948	0.552	Neg

The models of Bird Flu classification are generated by using C4.5 algorithm depicted in Figure 2, 3, 4 and 5 (see appendix).

Figure 2. shows the Bird Flu model created by Decision Tree algorithm The left subtree of Bird Flu model is shown in Figure 3. This model of non-serious cases since no poultry was destroyed in the farms. However, there were some positive cases indicated in this model.

Figure 4 showed the subtree(A) of this model which represents the conditions that less than 97the number of poultry were destroyed in these areas. There were many positive cases occurred in rainy season. However, in summer, there was only the negative case.

Last, Figure 5 shows the sub tree represents numbers of poultry were destroyed more than 97 cases in those infected areas. In this sub tree (B), there are many positive cases of AI infections than negative cases.

From the classification model, it shows that most factors affecting the Bird Flu pandemic including poultry farm region is central part of Thailand and month is October and humidity is more than 60 % and temperature is more than 27.7 Celsius. No pandemic occurs in summer time.

Further another factors contribute to Bird Flu pandemic include the Winter Season (November and December) and Humidity <=75% and poultry farms located in the central part of Thailand.

5. Conclusions

From the above results, we conclude that Decision Tree (C4.5) algorithm is enable to generate a classification model. This model might be used to predict the infected site of bird flu in poultry farms of Thailand. Further, factors relate to Bird Flu pandemic in farms are proposed. These factors include central part of Thailand, winter month, humidity = 60-75%, temperature>27.7 degree Celsius.

This model might be helpful in predicting the site of Bird Flu infection. The government, health related agency and poultry farms might use this model to prevent and watch for Bird Flu pandemic in Thailand.

6. Future Work

We might use Decision Tree model from this project to develop the surveillance warning system of the Bird Flu pandemic in the future. Moreover, other classification algorithms might be used to build models and evaluate their classification efficiency.

7. References

[1] "Decision Tree Learning" Available at http://en.wikipedia.org/wiki/Decision_Tree_Learning,.

[2] K.Abbas, A.R.Mikler, R.Gatti , "Temporal analysis of infectious disease : influenza". In *ACM Symposium on Applied Computing P267-271, 2005.*

[3] K. Abbas , A.R. Mikler , Amir Ramezani, Sheena Menezes. "Computational Epidemiology: Bayesian Disease Surveillance". University of North Texas *In Proc. of the International Conference on Bioinformatics and its Applications*, FL, USA, December 2004.

[4] G. F. Cooper , ET AL." Bayesian Biosurveillance of Disease Outbreaks." *In Conference on Uncertainty in Artificial Intelligence* 2004.

[5] J. Li, Q. Ren ,Jianqin Yin and Xi Chen . "Study On Transmission Model of Avian Influenza" .*Proceeding of 2004 International Conference on Information Acquisition*

[6] Ranjit Kumar Upadhyay , Nitu Kumari, V.Sree Hari Rao , "Modeling the spread of bird flu and predicting outbreak diversity". *In press Non linear Annual Real world Application* 2007

[7]Paola Sebastiani , Kenneth D Mandl , Peter Szolovits, Issac s Konane , Marco F Ramoni , "A Bayesian Dynamic Model for Influenza Surveillance" . *Journal of the American Statistical Association* 2005

[8] C4.5 . Available at *http://en.wikipedia.org/wiki/C4.5_Algorithm*

[9] E.Hak , ETAL . "Modeling the health-economic impact of the next influenza pandemic in the Netherlands" Available *at www.Sciencedirect.com*

Appendix

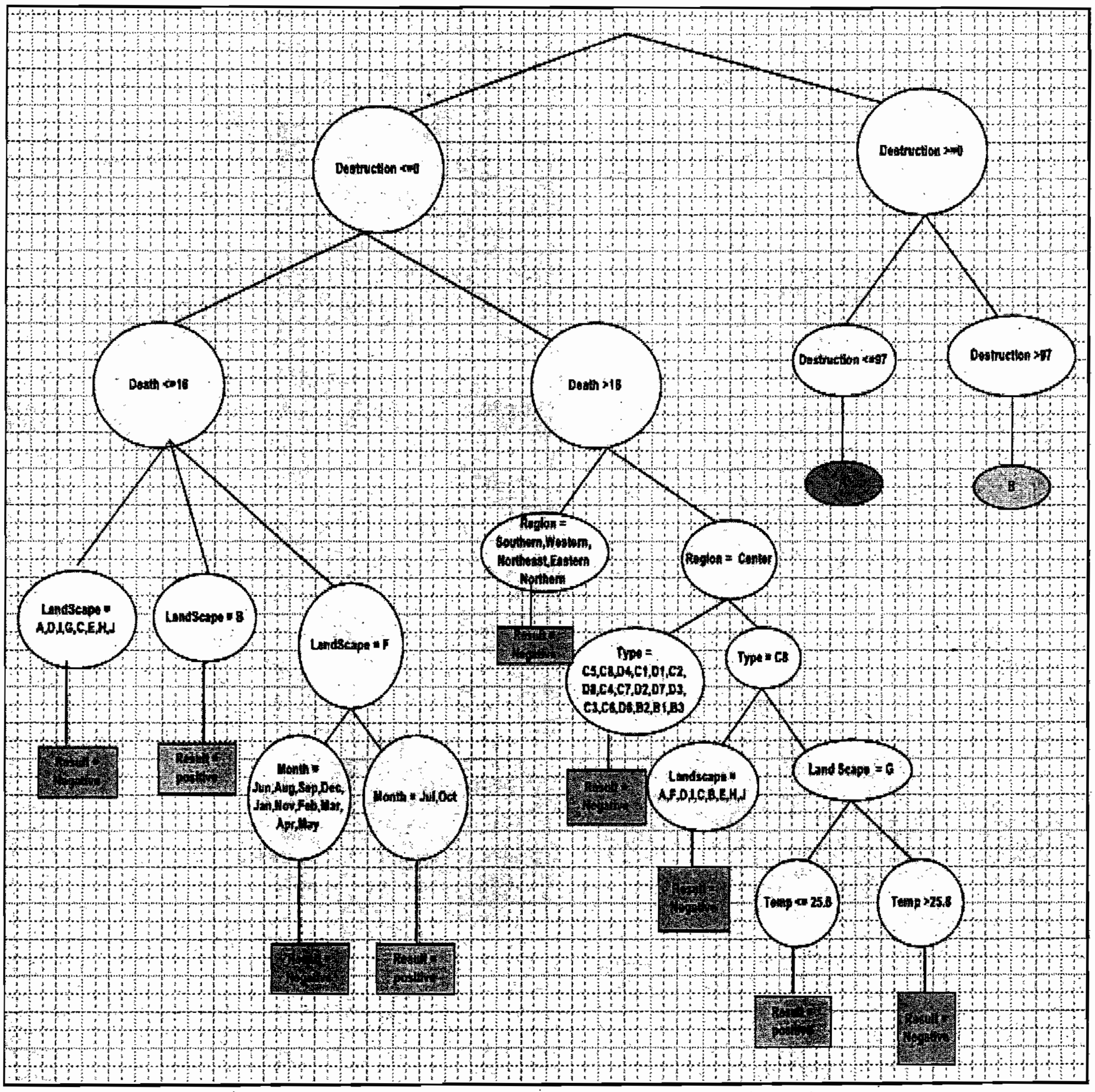

Figure 2. The final model created by using Decision Tree (C4.5)

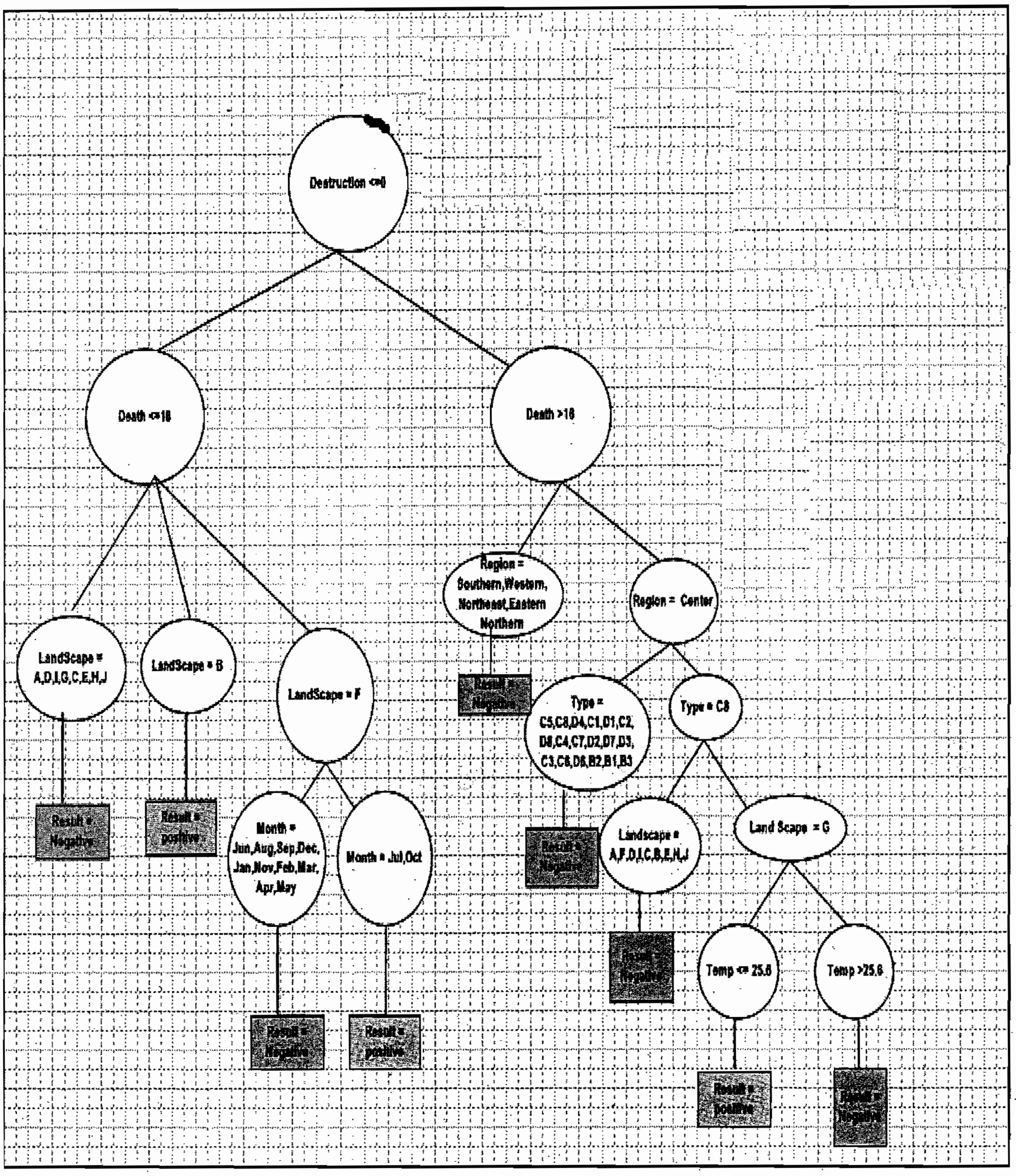

Figure 3. The model created by Decision Tree (left subtree of figure2: no poultries were destroyed)

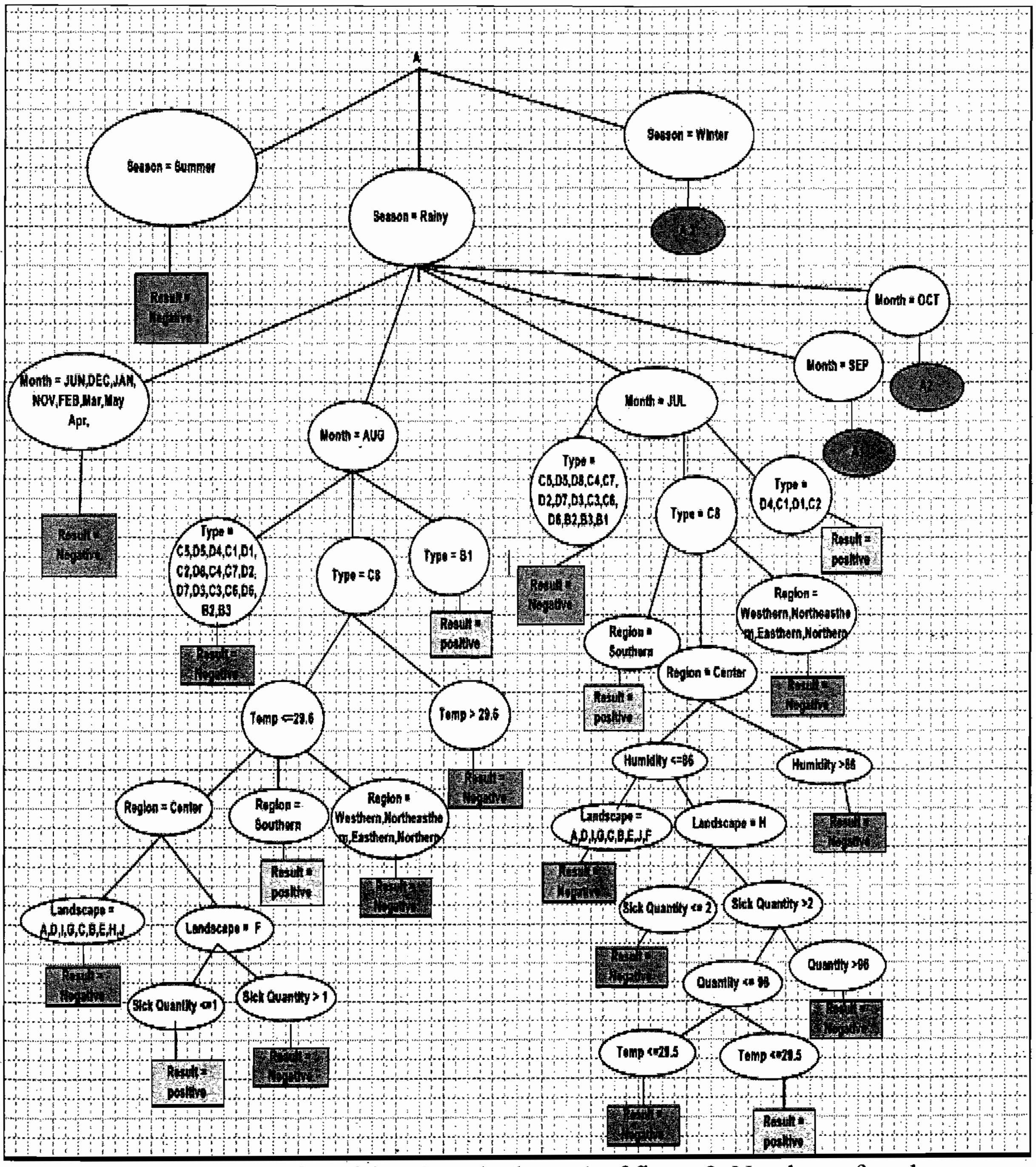

Figure 4. The sub model of Decision Tree (subtree A of figure 2: Numbers of poultry were destroyed from 0 to 97)

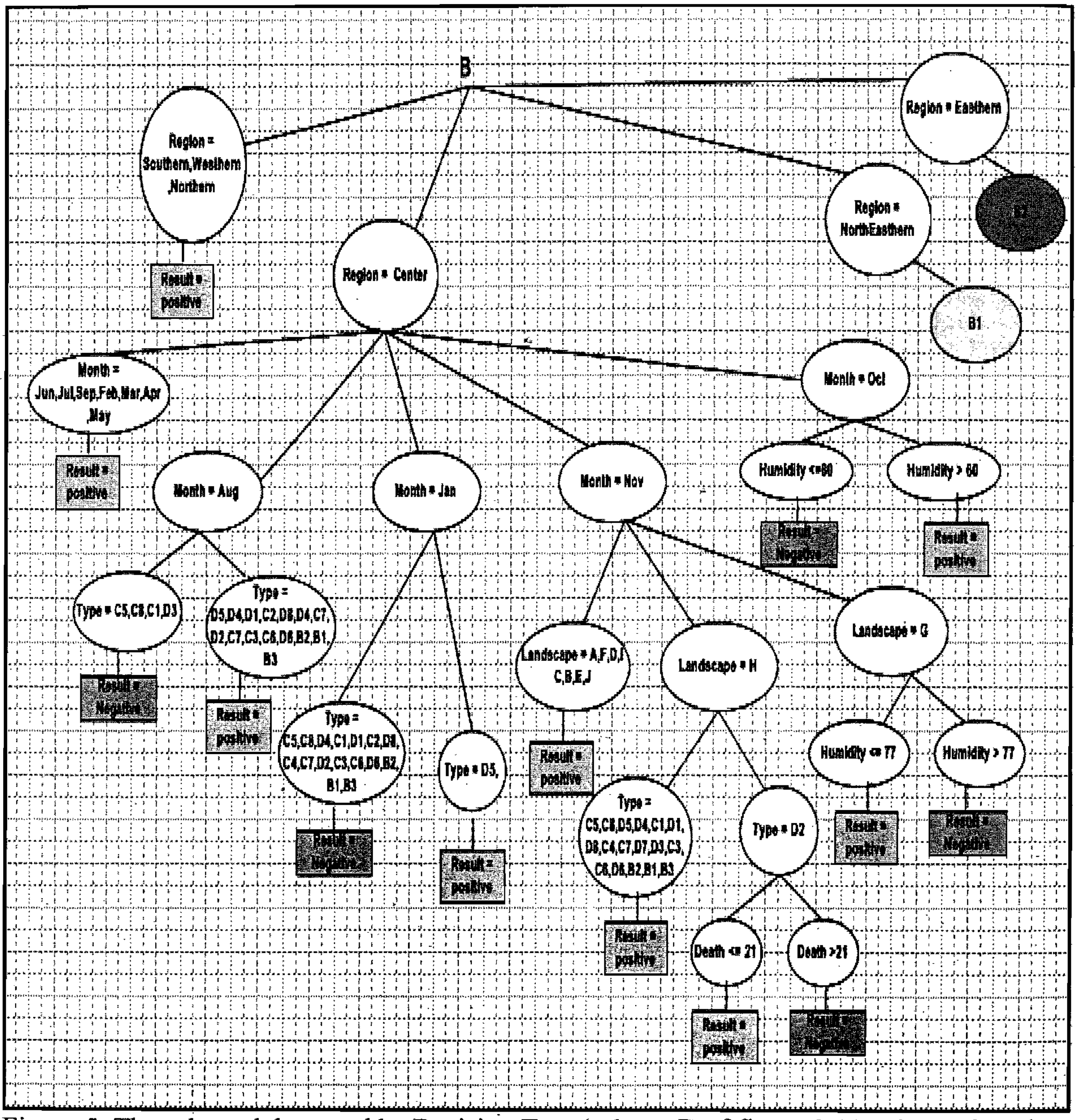

Figure 5. The sub model created by Decision Tree (subtree B of figure 2: Numbers of poultry were destroyed > 97)