DESIGN AND IMPLEMENTATION OF HEART DISEASE PREDICTION USING DECISION LEVEL FUSION MODEL

(A CASE STUDY OF UNIVERSITY OF ILORIN TEACHING HOSPITAL)

BY

ADETONA ABEL

HND/23/COM/FT/0530

THE RESEARCH PROJECT SUBMITTED TO THE
DEPARTMENT OF COMPUTER SCIENCE,
INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY
KWARA STATE POLYTECHNIC, ILORIN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF HIGHER NATIONAL DIPLOMA (HND) IN
COMPUTER SCIENCE

May, 2025

CERTIFICATION

This is to certify that this project was carried out by ADETONA ABEL with matriculation number **HND/23/COM/FT/0530** in the department of Computer Science Institute of Infor mation and Communication Technology, Kwara State Polytechnic, Ilorin.

-- ----------------------

MR ABIODUN E.T. Date

(Project Supervisor)

--------------------------------- ---------------------

MR.OYEDEPO Date

(*Head of Department*)

------------------------------------- ---------------------

External Supervisor Date

DEDICATION

I dedicate this project to Almighty GOD who's protect me, guide me through this work a b ig thanks to Almighty God.

ACKNOWLEDGMENT

With a grateful hearth, I return all glory to the most Merciful and Mighty Lord for the success of this Project.

My profound appreciation goes to my supervisor in person of MR ABIODUN E .**T** for his enormous contribution, support and guidance from the beginning to the end of this research work, I pray for her growth in all her aspirations in life.

Special thanks to HOD of Computer Science department, (MR OYEDEPO), and the member of staffs in the department of computer science for their academic support.

ABSTRACT

Artificial Intelligence is currently a popular theme in health care technology predictions. Machine learning is an artificial intelligence (AI) implementation that automatically learns and builds processes from experience. The leading cause of death worldwide at present is cardiovascular disease. The rate of death could be minimized by detecting the risk early. A lot of machine learning models have been developed to predict heart disease. We also introduced a fusion model to produce a better performance than the existing models. In this study, the proposed method analyzes two decision-making fusion models using five and ten-fold cross-validation to estimate the existence and absence of heart disease. The Jupyter Notebook, Scikit-learn, Tensorflow, and Keras were used as implementation tools. Three machine learning algorithms have been used here: the deep neural network (DNN), logistic regression(LR), and decision tree(DT). The decision tree and the logistic regression were merged separately with the deep neural network to form two fusion models (DNN+DT) and (DNN+LR), model-1 and mode-2. Five performance measurements have been used to compare model performance: accuracy, recall accuracy, f1-score, and AUC score. A significant improvement was found in performance parameters after fusing the algorithms in this work. The fusion occurred at the decision level by adding the decision scores of two algorithms. The main target is to enhance the fused model's performance by combining the individual model's decision for better classification. For both model-1 and 2, the accuracy has increased. Model-2 has obtained 87.12% classification accuracy for 10-fold cross-validation, which is the highest accuracy.

TABLE OF CONTENTS

CHAPTER FOUR: DESIGN, IMPLEMENTATION AND DOCUMENTATION

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATION

CHAPTER ONE

GENERAL INTRODUCTION

1.1 Background to the study

The number one worldwide cause of death is cardiovascular diseases (CVDs). CVDs are a category of problems of the cardiac and blood vessels, including stroke, heart attack, rheumatic heart disease, coronary artery disease, and other illnesses. Due to strokes and cardiac attacks,there are four out of five deaths, and people below the age of 70 die prematurely because of it, which is one-third of these deaths (Ayatollahi, et al. 2019). There are several reasons for CVD: unbalanced diet, smoking, stress, alcohol, fast foods, and inactive lifestyles. A study surveyed in 2016 that over 17 million individuals' deaths are because of cardiovascular disease by the world health organization, which accounts for over 30 percent of deaths worldwide. The same survey found that the number of mortality in under privileged and medium-income countries is more than 70%. The good news is that heart diseases can be prevented by avoiding some critical factors, such as poor diet habits and insufficient physical exercise. To control their general state of health and avoid sudden cardiac failure, prompt detection and predictive mechanisms are needed for people who are at risk of high cardiovascular disease. Speaking about predicting heart disease, one of the well-known prediction is machine learning. AI showed promising outcomes in healthcare. In the Journal of Clinical Analysis (Wang & Summers, 2012) in a 2012 study; it was reported that machine learning plays a significant role in automatically detecting intricate patterns in radiology applications, and it helped radiologists make smart decisions.

Kourou, et al. (2015), the researchers showed that machine learning is essential to improve our understanding of cancer progression. It also has a significant improvement in their accuracy and efficiency in decision making. Clinical diagnosis is a diagnostic task whereby evaluating the qualities of a variety of features; a doctor tries to classify disease. Traditional approaches to treat heart disease, including ECG, blood pressure, level of cholesterol, etc., are costly and require a lot of time. So, to drop the death rate. It is necessary to design a heart disease diagnosis system that is computerized. The number of physicians is low compared to patients. That is why it is essential to develop a medical diagnosis system to identify heart disease that is built on machine learning. It gives more precise results than traditional ways and reduces cost (Karayılan&Kılıç, 2017). Researchers are continually doing their hardest to identify more reliable smart models forsuccessful treatment of this disorder, and a variety of smart models based on professional ML methods have been established with the passing of time. To make the medical diagnosis easy and cheap, we choose to build a fusion model to classify disease effectively. Here, a decision level fusion approach has been presented for heart disease disorder. Decision level fusion has implemented using the summation of scores of separate models (Matin, et al. 2017).The decision score of the separate models has been taken from the trained models. As the fusion occurs at the decision, so it is referred to as decision level fusion.

The supervised algorithms- decision tree and logistic regression are fused with a deep neural network separately to develop a better fusion model that can tell the existence and non-existence of heart disease by analyzing the data. This work contributes by making a fusion modelbesides different machine learning algorithms to achieve better accuracy than them. And in our work, there was a notable improvement in the accuracy after fusion. This work is presented as follows. The concern for the issue of heart disease is discussed in the first segment. The second segment contains the previous work. Materials and methodology are presented in the third segment. In part four, the paper explains the experimental out comes of our suggested architecture for identifying heart disease and also a comparative analysis with previouswork. After that, we have ended with a conclusion and prospects.

1.2Statement of the problem

A lot of work has been performed on the classification of heart disease. The goal of all researchers was to improve accuracy. Various algorithms have been applied to the dataset to observe which model performs best. Now this work has been trying to build a fusion model for effectively classifying disease efficiently. Different algorithms are now making using fuzzy logic or merging algorithms to create a new one. This type of compu tational intelligence techniques plays an important role in medical diagnoses. Hopefully, these techniques make the physician's works easier. In the future, the same techniques c an also be applied to other disease prediction and also some other intelligence technique s will be applied to predict the coronary artery heart disease. The diversity of resources w ill deliver improved performance in knowledge extraction and a clear understanding of th e measuring and collecting data problems. The real-time application will be updated and more features will be added and a real-time website will be designed for the prediction of heart disease. With the help two artificial intelligence Deep Neural Network- Decision Tre e (DNN-DT) and Deep Neural Network- Logistic Regression (DNN-LR), the project develo p a system for heart disease prediction system.

1.3 Aim and objectives

The aim of this research work is to develop a comparative analysis of two intelligenc e based decision level fusion model for heart disease prediction using Deep Neural N etwork- Decision Tree and Deep Neural Network- Logistic Regression. The objectives of this project are to:-

i. Develop a comparative analysis system of two intelligence based decision fusion model

ii. Predict heart disease develop in (i) using Deep Neural Network- Decision Tr ee and Deep Neural Network- Logistic Regression

iii. Evaluate the performance of Deep Neural Network- Decision Tree and Dee p Neural Network- Logistic Regression in prediction system

1.4 Significant of the study

The recent advancements in AI techniques Deep Neural Network- Decision Tree and Deep Neural Network- Logistic Regression have played an important rol e in biomedical sciences providing a handy role in diagnosis and monitoring of va rious diseases. The comparative analysis of two intelligence-based decision level fusion model has shown a robust system for the prediction of heart disease.

1.5 Scope and limitation

A comparative study of computational intelligence techniques in this study Deep Neural Network- Decision Tree and Deep Neural Network- Logistic Regression (LR, DNN, DT,) a nd the developed system used the same dataset but the implementation techniques are different so we take the direct experimental result from the work.

1.6 Organization of Report

Chapter one contains the background to the project, statement of the problem, the ai m and objectives, significance of the study, justification of the study, scope of the study, definition of technical terms and organization of the report. Chapter two entails the litera ture review that talks about previous and related work of proposed system.

More so, chapter three examines the method of the proposed system, the existing s ystem which includes analysis of existing system, problems of the existing system, and t he description and benefit of proposed system. Chapter Four, explains the implementatio n and documentation of the system which contains design, implementation techniques, documentation, hardware and software requirements. Lastly, chapter five includes the su mmary, conclusions and recommendations.

CHAPTER TWO

LITERATURE REVIEW

1.1 Review of Related Past Works

Christensen, et al. (2018) developed a systemfor different types of disease predict ion, four different machine learning approaches were analyzed:LSTM, which is a kind of the recurrent neural network, XGBoost (XGB), random forest (RF), and Logistic regression (LR). Here, XGBoost is performing better than LSTM.

Krishnaiah, et al. (2016) studied a fuzzy k-NN classifier was used, and the perfor mance of the fuzzy k-NN classifier was way better than k-NN classifier. To remove the da ta'suncertainty, Fuzzy k-NN was used, and it provides higher accuracy than the k-NN clas sifier.

Tougui, et al. (2020) studied six machine learning algorithms were applied to Clev eland dataset using six data mining tools, and then the result of those algorithms was c ompared to each other. Based on the data examination and the effects of the output dat a taken, the best output was given by ANN which was implemented in Matlab,and the be st performing tool was Matlab.

Dwivedi, (2018) study several machine learning algorithms were used for medical diagnosis. These methods have been validated by tenfold cross-validation. Logistic regr ession gives an accuracy of 85% that was highest with an F1 score of 79%. And ANN wit h an accuracy of 84% with an F1 score of 84%. Among other things, those two algorithm s worked better.

Birjais, et al. (2019), work on a system that diabetics dataset was used for classifi cation. It was predicted if a person has diabetes or not. The prediction was made on test data with three-fold cross-validation. In this work, Gradient Boosting offers a maximum a ccuracy of 86%, which is higher than Naive Bayes and Logistic Regression. Both Naive B ayes and Logistic Regression showed an accuracy of 77% and 79%, respectively.

Karayılan&Kılıç, (2017) predicted heart disease, and a multilayer perceptron was implemented in the suggested architecture. The architecture of the neural network contains 13 features derived from the Cleveland dataset. The proposed system gives 95% accuracy, and different accuracy has been found with the variation of the number ofhidden layers. The improvement in our work compared those that we have used the fusion model to identify the disorder, which results in improved outcome.

Bharati &Podder, (2019) focused on creating a hybrid model to diagnose lung disease from X-ray images. CNN, Vgg16, and other techniques have been applied to predict lungdisease. As CNN gave a poor performance, that's why a hybrid model was implemented for better results. The hybrid model VDSNet outperforms current methods interms of the evaluation metrics.

1.2Decision Tree (DT)

A decision tree is a tree-like structure consisting of branches, nodes, and leaf nodes. It is a branching graph that functions like a splitting rule for each particular attribute. Each feature is treated as a branching node. These nodes build a rule, and, based on the rule, values are grouped into different classes. In the decision tree, the leaf decides some decisions at ending, and the topmost is theroot, which partitions the tree based on attribute value.Building A DT is easy and simple, and the results arepredicted more accurately.DT is sensitive to overfitting. It happens when themodel is very good at identifying trained data but givespoor performance for test data. It becomes over skilledfor training data by having minimal impurity in the leafnode. That's why pre-pruning is needed to minimizethe number of leaf nodes that are not that importantfor model building. It gives better accuracy for prediction. Information gain is another important criterion,and attributes with the highest information gain splitfirst. For that method, features with the lowest entropyare selected for splitting (Naz& Ahuja 2020). In the proposed design Gini was used as criterion, maximum dept of the tree was 8,and minimal leaf size was 10. Pre-pruning was done toget the best parameters. The Gini index and entropy aresplit criteria for a decision tree, and they are arranged by applying (3) and (4).

Gini: $Gin(E)=1-f(x)=1-\int nj=1Pj2$ (3)

Entropy: $(E)=1-f(x)=-\int nj=1Pj\log 2Pj$ (4)

Here Pj is the percentage of a class in a node.

2.3 Deep Neural Network (DNN)

An ANN has more than one layer, then is named a deep neural network (DNN). Lik e humans, a deep learning (DL) system can teach itself and learn through several hidden layers. DL is a model that is based on a multilayer feed-forward perceptron and uses bac k-propagation to train with stochastic gradient descent. There are four layers with nodes and neurons for the designed network,which has a uni-direction. It has two hidden layers, and there is a single way connection between every node and the next node. For training, stochastic gradient descent was used with back-propagation. It has been suggested as a very useful tool in different medical sectors for decisionmaking. Proposed Architecture o f Deep Neural Network In this proposed approach, an input layer was used for data entra nce, an output layer for predicting out- put(presence/absence), and two hidden layers. On e of the most challenging tasks is model optimization for machine learning implementat ion. Model optimization reduces the test error; however, deep learning tunes theparamete rs outside of the model but has a significant influence on results and classification. The mother of all hyperparameters is the learning rate. The speed of learning of the model de pends on the learning rate. The number of hidden units is vital as it regulates the model's representational capacity. The number of first and second hidden layers in our model is 10 and 8, respectively. L2 regularization was used to block overfitting. Bias, which is a constant, was used to help t he model in a way that can fit best for the given data. Batch size ten was used.

2.4 Logistic Regression (LR)

LR is a supervised machine learning technique with continuous/discrete predictors. LR is generally used for binary classification. It is a statistical model that represents the relatio n between the binary dependent variable’s logit transformation and independent variable s (one or more than one) by determining the best fitted linear model. This model is a simple prediction a pproach compared to other non-parametric models of machine learning with baseline ac curacy scores provided by the model (Cox, 2008)

2.5 Computational Intelligence Techniques

Computational intelligence technique is a collection of computational models. Usi ng it many kinds of problems can be solved easily which are difficult to solve using conv entional computational algorithms. There are different applications of computational int elligence in the field of medicine and pathology (Rubio, et al. 2015; Shahid & Singh, 201 9; Tizhoosh&Pantanowitz, 2018). The LR, SVM, DNN, DT, NB, RF, and K-NN algorithms w hich are used in this comparative study are explained in details. Another interesting meth od is reinforcement learning. Reinforcement Learning works better for classification prob lems for imbalanced dataset. For classification problems, reinforcement learning has ser ved in removing noisy data and learning better features, which made a great improveme nt in classification performance. But this technique works better for a large amount of d ataset and image dataset. As the dataset size of Cleve and Stat log are not so big and all the attributes are numeric, hence reinforcement technique is not better suited for this stu dy

2.5.1 Common Techniques for Data Mining

Naïve Bayes, Decision trees, Genetic algorithms, and Neural Networks are few of s everal techniques that have been proposed and used for mining data in recent times. Naï ve Bayes, popularly known as Bayesian Network and decision trees are common classifi ers applied in many works. For instance, Edin and Mirza (2012), applied decision trees m odeled from Naïve Bayes on preoperative assessment data to predict students’ performa nce in a summer course. Genetic algorithms have been applied to predict their students' final grade by (Minaei-Bidgoli& Punch., 2013). However, neural network deems less suitable for data mining purpose due to lack of comprehensibility because such models are built on black-box mechanisms. Genetic Algorithm and Neural Network have been hybridized in (Yao 2009), and neural evolutionary programming is proposed in (Martínez., 2016). Review of these operational methods is given herein.

i. **Naïve Bayes :** Naïve Bayes is a data mining classification technique that applies Bayes' theorem with strong independence assumptions between the features using simple probability. This classifier is scalable in that it requires a number of parameters linear with number of attributes in problem domain and maximum likelihood training only takes a linear time because evaluation involves a closed-form expression rather than by expensive iterative approximation as used by many other classifiers (Martínez., et al 2006).

To classify an item x_i as an instance of X using Naïve Bayes, the classification model computes a posterior likelihood of x_i as:

$$P(C_i X) = P(C_i) \sum_{i=1}^{n} P(x_i|C_i) \dots \dots \dots (1)$$

Where C_i is a classifier label and $P(C_i|X)$ is the priori probability that an instance of X is tagged (labeled) in C_i .

Naïve Bayes is widely used for learning data relationships unlike other complex classification techniques (John and Langley 2005). Despite its simplicity, computational efficiency and performance for real-life problems, results from decision tree are not that accurate making its potential for probabilistic modeling unexploited.

ii. **Decision Trees:** Decision tree is a support tool with tree-like model used in taking decisions and showing relative consequences, chance of event outcomes, resource costs, and utility. This approach, commonly used in operation research helps to identify a strategy most likely to reach a goal. In a flowchart-like tree structure, the internal nodes denoted by rectangles are used to test values of expressions on given attributes while leaf nodes denoted by ovals depict the class label an entity can be categorized.

Decision trees can adopt classification or regression tree analyses. In the classific ation tree analysis, predicted outcome is a class to which data belongs while in regressio n analysis outcome can be a real number. Most decision tree algorithms such as ID3, C4. 5, and CART adopt top-bottom, and divide-conquer procedures to learn from training dat aset (Kotsiantis 2005). These algorithms starts by assuming an empty tree then split on next best attribute, and hold on to recursion on each leaf. Choosing the best attribute up on which splitting is done is a great challenge. This could be addressed by quantifying t he predictive feature of an attribute since outcome at current nodes depends on this. In C ART, variance reduction is often employed in cases with continuous target variable (Yi-H orng 2015).

Other recent algorithms include Chi-squared Automatic Interaction Detector which performs multi-level splits when computing classification trees (Satya et al., 2011), multi variate adaptive regression splines extends decision trees to handle numerical data bette r. The regression technique is a non-parametric extension of linear models that automati cally models interactions between variables. In general, there might be multiple independ ent variables but their relationship to dependent variable(s) is unclear by manual analysi s and neither is it visible in plot. Hence, regression techniques are used to discover non-li near relationships in multiple variables.

Decision trees represent rules in format that can be easily read, understood and interpret ed by users. Over time, decision trees had been built with conditional inference. Conditio nal inferential is a statistical technique that takes non-parametric tests as splitting criteri a for multiple testing to achieve unbiased prediction without pruning nor over-fitting. ID3 and C4.5 were older techniques used to generate decision trees from a dataset. C4.5 has realized a number of improvements over its successor ID3. An example is its ability to ha ndle both discrete and continuous attributes.

iii. **Neural Network:** Neural Network is an artificial intelligent technique that hosts a gro up of interconnected artificial neurons to mimic the properties of biological neurons in human. The technique follows analog and parallel computing system made up of simple processing elements that communicate through a rich set of interconnections with varying contributory weights. Over a period of time, neural network has been used to solve pattern recognition problems such as text, image and medical bioinformatics. Kim et al., (2016) a predictive model based on neural network was applied to estimate the compressive strength of concrete mixture from constituent's proportion and the prediction accuracy was improved with an iteration method.

Since neural network is an adaptive artificial intelligent technique with self-tuning used that adjusts its structure during learning process, its usage in students' academic performance prediction can be described as being probabilistic. Probabilistic neural network, implemented based on Parzen non-parametric probability density function estimation and Bayes classification, use feed-forward network to effectively solve prediction problems. The approach was found to build its architecture and train the network in fewer seconds, thence provides probabilistic viewpoints from all input with deterministic classification results. The Probabilistic network usually adopts Radial Basis Function to map any input pattern for classification.

Neural networks have shown success in areas of performance prediction, estimation and approximation, and pattern recognition problems. Probabilistic network model was proven to be more time efficient when compared to conventional back-propagation based models. The network type is recognized as a way to solving real-time classification problems as observed. However, it works better if a desired output is expressed as one of several pre-defined classes. Finally, NN and some statistical methods, like decision trees built on conditional inference, are deemed less suitable for data mining purposes.

CHAPTER THREE

METHODOLOGY AND ANALYSIS OF THE EXISTING SYSTEM

3.1 Research Methodology

Two decision level fusion models were constructed by applying an artificial neura l network, logistic regression, and decision tree. Three algorithms were applied individuall y on the same dataset, and then the decision scores of the DNN and DT algorithm were c ombined as well asthe scores of DNN and LR.

The fusion models Deep Neural Network- Decision Tree and Deep Neural Network-Logist ic Regression(DNN+DT) and (DNN+LR) gave a better result than individual achievement.

The two Artificial Intelligence Based Techniques

i. Deep Neural Network- Decision Tree(DNN+DT)

ii. Deep Neural Network-Logistic Regression(DNN+LR)

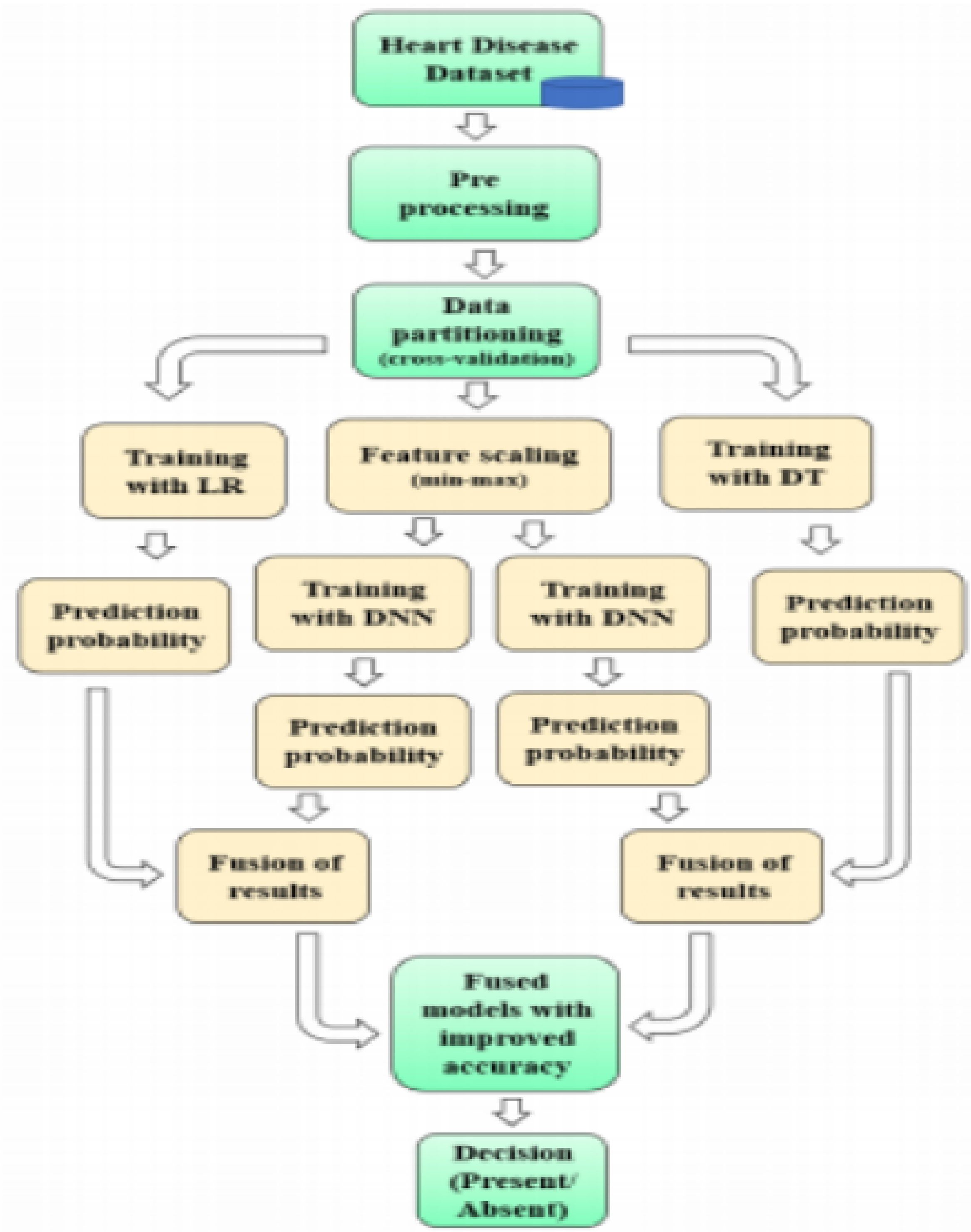

Figure 3.1: Proposed Architecture Model of Deep Neural Network- Decision Tree and Dee p Neural Network- Logistic Regression (Karayılan, 2017).