Machine Learning To Predict Child Growth

Conventional Research Project
Supervisor: Michael Kirley
Subject Code: COMP90055

Deshna Jain
Student ID: 936309
School of Engineering
The University of Melbourne
Melbourne, Australia
deshna@student.unimelb.edu.au
Credit Points: 25

Logesh Chinsu Palani
Student ID: 796735
School of Engineering
The University of Melbourne
Melbourne, Australia
lchinsu@student.unimelb.edu.au
Credit Points: 25

Tarun Dev Thalakunte Rajappa
Student ID: 934175
School of Engineering
The University of Melbourne
Melbourne, Australia
tthalakunte@student.unimelb.edu.au
Credit Points: 25

***Abstract*—One in every four children in Vietnam suffers from stunting. Stunting is dependent on multiple clinical and social economic factors. High performing and accurate machine learning models on this data can drive the success of identifying the determinant factors for stunting in a child. In this study, various imputation methods, twelve feature selection methods and three machine learning models are evaluated based on their performance in predicting stunting. Publicly available implementations are used in the study and the parameters are reported. From the results it is concluded that univariate feature selection methods when used with weighted knn and smote imputed dataset gives the highest performance and statistical tests prove that neural networks is more significant than the other two models.**

***Index Terms*—Stunting, feature selection, machine learning.**

I. Introduction

Undernutrition is a major public health problem it leads to premature morbidity and mortality in early childhood and accounts for 45% of deaths under five-year-old (Lartey, 2015). It refers to the absolute or relative deficiency of one or more nutrients. In 2016 it was estimated that globally 151 billion children below 5 years of age had stunting or chronic undernutrition. Of this total, 57% of the children reside in Asian countries. A child is said to be stunted when standard deviation of low-height-for-age is lesser than -2. The prevalence of stunting in South Eastern counties (Vietnam) is found to be 25.8% of the population below five years (Lartey, 2015). Childhood undernutrition increases the risk of cognitive development, death and health status over the years. These children are likely to have recurring sickness and faltering growth. Critical associations between stunting at age of 2 years and long-term consequences are reported by the Lancet series on maternal and child undernutrition (Lartey, 2015). This is a growing cause of concern for the countries as it hinders the socioeconomic growth of the country but also impedes the mental and physical growth of children. Even though there has been a considerate decrease in the stunting over years it is still large compared to the global rate.

There are multiple factors which cause undernutrition during childhood such as poor health, mother's education, height at birth, social economic factors, etc. These factors vary for different countries and it is critical to identify the optimal attribute that affects the stunting in a child. In this research we are using real-time clinical data collected by Doherty Institute from Vietnam about parents, infants and social-economic factors.

There are three aspects of novelty in this paper. Firstly, a comparison study of different pre-processing methods for the high dimensional real-time medical dataset. Secondly, to identify the determinants of stunting using a comparison study of various univariate and bivariate feature selection methods. And finally comparing the different machine learning models that best classify stunting.

In this study, we help identify candidate children for early stunting intervention by using data that monitors women through pregnancy, an infant at birth, socioeconomic factors and father's clinical data. Doing this targets the children that are at risk at a very tender and critical age of development.

II. Literature Review

Most of the previous studies which determine the prevalent factors for stunting focus on linear or logistic regression for predicting child growth. According to researchers the most affecting factors for stunting in Bangladesh (South Asian country) are household wealth, food insecurity, mass media exposure, age of the child, size at birth, etc (Sarma et al., 2017). Most of these factors can be correlated. It is commonly noticed that households which were moderately or mildly food insecure have a higher number of stunted children than otherwise (Ali et al., 2013). Studies conducted in the rural parts of South Asian countries suggest that the risk factors of stunting increase with lack of knowledge about the initiation of complementary feeding, nutrition and dietary variety (Ahmed, Ahmed, Roy, Alam, & Hossain, 2012)(Menon, Bamezai, Subandoro, Ayoya, & Aguayo, 2015). These factors stated in these papers are determined mostly by surveys, human intuitions, and facts. The factors identified here are restricted to a geographical location due to data source location and the time period over which it was collected. When data from

different geographical location and over separate timeline the features identified could be explicitly different.

In (Cruz et al., 2017) the authors use binary logistic regression to determine the factors that affect stunting in the central region of Mozambique. Each of the factors is compared with the stunting label and backward elimination is performed and the collinearity is tested. Similarly, in (Prendergast & Humphrey, 2014) the determinant factors identified are the nutritional status of the mother, feeding practices, sanitation and access to healthcare. These papers only consider the factors that affect stunting directly and not their influence on it collectively. Also, it does not provide any proof of how well the determinant features can predict stunting.

There are very scanty research papers which use Machine Learning to identify the determinants for stunting despite the availability of high computation in the era of big data. One such study is (Khare, Kavyashree, Gupta, & Jyotishi, 2017) where the author compares the features identified by literature review and by machine learning techniques. In this study Weka is used as a tool there is a comparison of the highest accuracy value that occurred with the top number of features. The authors do not explain the sensitivity and specificity of the test results. The indicating factors selected using entropy and information gain are not checked for their independence and are only chosen based on the feature ranking.

Paper (Markos, 2014) exhibits the usage of machine learning technique to find determinants that identify undernutrition of children before 5 years of age. These researchers also use Weka for feature selection. The researchers do not display the results that were obtained from other methods such as J48 decision tree and Nave Bayes. The paper is focussed on the PART rule induction. Data used in this study is converted from continuous value to discrete value, doing so a lot of essential data is lost. The paper does not discuss in detail about the feature selection methods used and compare the different methods.

In the study(Zhang et al., 2009) to compare the logistic regression and machine learning methods in predicting obesity at childhood the authors use many machine learning models such as SVM and neural nets but the data has only 3% of data samples of obesity. There is no preprocessing done to overcome the class imbalance. This data has important consequences on the learning process of the classifier. Therefore the classifier have a poor accuracy for the minority class and further increases bias towards classification into majority classes.

With a large number of features present in the medical dataset, there are a few studies which use machine learning techniques to identify prominent features and compare the different classifiers. Machine learning is extensively used in everyday life but in the field of medical and health, the usage is limited. The ability of machine learning models to deal with high dimensions of attributes, ease of use and categorizing thousands of data makes it suitable for this research. This research applies machine learning and feature selection techniques to identify the optimal determinant factors that affect the stuntedness of a child, predicting future interventions based on the important factors and compare the diagnostic property. We also compare the performance of the three different machine learning models and feature selection techniques.

III. Methodology

A. Data Preprocessing

In real-world settings, nearly all the surveys and census suffers from incomplete data. The given dataset from Doherty Institute is the result of long going study related to stunting in Vietnamese children and the corresponding survey tend to have a high percentage of missing values. As the classifiers require a complete dataset, It is vital to handle missing data. Many approaches to handle the missing value are considered including listwise data deletion, KNN imputation and weighted KNN with attribute association imputation.

1) Case deletion (CD): One of the common strategies to deal with missing data is to simply ignore the instances or attributes with a high level of missing data. The given dataset contains 1168 instances with 165 independent feature. As every instance contains a missing value, deletion of rows with missing instances will yield zero records. Hence, firstly 15 attributes which had more than 60 percent of missing data are removed and consecutively the instances with missing data are removed yielding 137 instances with 150 independent features. However, removing data from the original dataset could introduce bias when the missing data are not missing randomly(Little & Rubin, 2010)(Rubin, 2011). Hence, the method of imputing data for the missing value is considered.

2) K nearest neighbours: In the study(Troyanskaya et al., 2001) a nearest neighbor based method for imputing missing value is provided which is widely used in the medical dataset. This imputation method selects instances which have similar attribute values to the instance with missing value. For each instance with a missing value, the method calculates the distance of all the other records using Euclidean distance and finds K closest instances. The missing value is then calculated as the weighted mean of the neighboring values weighted by the Euclidean distance of the neighbors. Although KNN imputation is widely used and easily implemented for high dimensional data, this method can be applied only to continuous data and cannot be applied to ordinal or nominal categorical data(Faisal & Tutz, 2017). Hence, for the given dataset, KNN imputation is applied only on the continuous data considering 5 nearest neighbors while imputing missing data.

3) Weighted KNN with Attribute association: For imputation of categorical data, the complex structure of the correlation between attributes has to be considered. (Faisal & Tutz, 2017) proposes a method that imputes missing data by calculating the distance which takes the correlation between the attribute of the missing data and other attributes into account. The highly associated attributes have more weight

and contribute strongly than other low weighted attributes towards computation of missing value. For each missing value Z_{is} in instance i and attribute s, the Euclidean distance d_{CatSel} is calculated. This distance takes the association among the attributes into account, and is calculated between the instance i and all other instances. The 10 nearest neighbors are chosen based on the Euclidean distances. The weights for all the selected neighbors are calculated using the Gaussian kernel density function. For each prospective categorical value for the missing value, weighted estimators are calculated by adding the kernel weights of all nearest neighbors. This contains the categorical value in attributes.

The imputed categorical value is given as the categorical value with the highest weighted estimator. Thus this method considers the Euclidean distance between the missing data instance and nearest neighbors as well as the correlation between the attribute of the missing data and other attributes and imputes a proper categorical value for the missing data. Since both the distance and correlation are measured by the similarity between instances and attributes, there is no bias added in the imputation process.

4) SMOTE: It is observed that in the given data only 14% of the training data is classified as stunted. Performing machine learning or feature selection on this data is unhelpful. Many studies have shown that having balanced data leads to better prediction performance. Therefore a sampling method is introduced to modify the imbalanced dataset to an equal distribution in the class label. Synthetic Minority Oversampling Technique (SMOTE) is a sampling method that is proven to be powerful, it is commonly used in machine learning model with high dimensional imbalanced medicine dataset.(Blagus & Lusa, 2015). This technique increases the number of minority class (ie. Stunting) by randomly generating instance from the nearest neighbor line of the minority label. The created instances are based on the feature set of the data and are similar to minority instances. In the study(Alghamdi et al., 2017) it is established that using SMOTE showed significant improvement in prediction and classification of incident diabetics.

5) One hot Encoding: The next preprocessing method is one-hot encoding. In this the categorical data with label data is converted into multiple columns. If the encoding is not done than the machine learning model assumes that the data has an order or hierarchy. Performing one-hot encoding allows a better presentation of categorical data especially in high cardinality categorical data (Cerda, Varoquaux, & Kgl, 2018) . Following the one-hot encoding four different datasets are created by pre-processing.

6) Scalar standardization: (Buitinck et al., 2013) proposes scalar standardisation that transforms the feature data to be centered around zero with unit variance. If a feature has a larger variance compared to other features, then it tend to have more influence in the estimator function. The estimator will not learn from the feature with comparatively low variance. To avoid such differences, scalar standardisation is applied on all features to bring them in a relatively similar scale. The machine learning algorithms used in this study also expects and assumes that all the feature data have similar distribution (Hall, 2016). After applying standardisation on train data (Buitinck et al., 2013) suggests that the mean and standard deviation used for each feature should also be applied to the respective feature test data.

B. Feature Selection

The success of machine learning on a medical dataset can be affected by many factors. One such dependent factors are the quality of the data. The information extracted from the large feature set data is redundant and irrelevant or when the data is noisy. This hinders and makes knowledge gain during training process more difficult. Applying machine learning on this noisy data will lead to overfitting thereby reduces the performance of the classification model on unseen data.

To overcome this issue feature selection methods is implemented. It selects a subset of available features which are relevant and has immediate effects such as decreasing the computation, memory storage requirements and improving the performance of the model and the result time comprehensibility(Guyon & Elisseeff, 2003). This is a part of preprocessing the data and preparing it for the machine learning models. It is necessary to preprocess as in higher dimension the data becomes sparse and affects the machine learning algorithm which is typically designed for low dimensionality. This phenomenon is referred to as the curse of dimensionality.

Curse of dimensionality can be tackled by reducing the dimensionality, the reduction has two components: feature selection and feature extraction. Feature extraction converts the original dimensional features in a linear or nonlinear combination of the original feature set. This cannot be used in this study as this removes the physical meaning of the feature and adds extra analytical time. To maintain the interpretability and readability of the model it is important to retain the original features. In this study, the aim is to find a correlation between the feature and the class label hence supervised feature selection methods are used. Feature selection methods are used to identify the most influential feature/reduce the dimensionality of the dataset, it is used either to improve accuracy or improve performance on such high dimensional data.

Feature selection can be classified into three different types: Filter method, wrapper method, and embedded method. Filter method selects the features independently of the learning model (Li et al., 2017). Wrapper method improves the quality of selecting features by relying on machine learning algorithms. In the embedded method, the whole feature selection relies on the learning module. In this study, we have used various filter feature selection algorithms. In this method, the features are ranked based on scoring criteria. Filter methods are of two types: univariate and multivariate. Univariate filter

methods scoring is dependent on the relevancy of the feature while ignoring the redundancy. Whereas multivariate is an aggregate of both redundancy and relevance. For the multivariate feature selection methods, the feature count parameter *n_selected_features* is set to 15. These algorithms are run with different machine learning algorithms to determine the best available features which predict with highest AUC in the machine learning model.

The different filter methods used in this project are following:

1) Mutual Information: In this multivariate feature selection method, the relevance of a feature is based on their measurement of correlation with the class variable. Shannon's theory states that entropy *H(c)* can determine the uncertainty of variable C. Mutual information can be defined as the measure of the amount of information one variable has about another variable. Searching for the best set of features globally is an NP-hard problem. To avoid the problem this method utilizes a heuristic sequential search where the features are added or removed one at a time. Good features, in reality, should not only be highly correlated to the class value but also not be correlated to other features. In this study for continuous variables estimation the number of neighbors parameter *n_neighbors* is set to 3.(Cover & Thomas, 2006) This feature selection method keeps in consideration both the feature relevance and feature redundancy.(Battiti, 1994)

2) Mutual Information Maximisation: This is also known as information gain it measures the correlation between features and a class variable. The assumption in this method is that stronger the correlation with the label indicates it will perform well in classification. The features are accessed individually, and the redundancy is completely ignored. Each feature is given a score and the top scoring features are added to the feature set (Lewis, 1992).

3) Minimum Redundancy Maximum Relevance: This multivariate method was proposed in 2005 by Ding and Peng. This feature selection method requires the discriminative features to be dissimilar to each other. The main intuition of this feature selection method is that when non-redundant features are selected then the new features selected will not be redundant to the already selected features. In the study (Zeng & Li, 2014) to identify feature detection for tumor it is concluded that MRMR is most robust to estimate feature redundancy in their experiments.

4) Joint Mutual Information: This is an alternative to MRMR and MI feature selection method. It works on increasing the complementary information that is shared between selected and unselected features provided the class label. In study(Liu & Motani, 2018). it is observed that this feature selection method is used to observe AUC of the medical dataset such as breast cancer.

5) Conditional Mutual Information Maximisation: It is an efficient and fast multivariate feature selection technique that is a variant of Conditional Mutual Information feature selection technique. The algorithm calculates the redundancy and relevance, then selects the features that maximize the MI with the class label with the condition to any selected feature. This attempts to achieve a balance between the independence of comparison and individual power. When tested in studies(Alzubi, Ramzan, Alzoubi, & Amira, 2018) it is observed to perform better than a few other feature selection methods.

6) Feature Importance: The feature importance is calculated using the Gini Index. It is a statistical measure that quantifies it the instance is rightly identifying instances of different classes. Lower the Gini index more relevant the feature is. This univariate feature selection method provides importance scores and can be easily implemented on high dimensional data. The parameters used are *n_estimators* is set to 10 and the criterion is Gini (Menze et al., 2009)(Geurts, Ernst, & Wehenkel, 2006).

7) L1 based feature selection: L1 based feature selection is a univariate feature selection method for reducing the dimensionality of the data. Linear Support Vector Classifier is used for classification, penalization is done using l1 norm. The error term penalty parameter *C* is set to 0.01, this parameter controls the number of features being selected. If the *C* value less then fewer features will be selected. The goal here is to identify top influential features, hence a very low *C* value is chosen. From the coefficient matrix, the features with the non zero coefficients are extracted and chosen for further process.

8) Generic Univariate Select: This method investigates each feature by comparing it to the class variable in terms of correlation. Based on the scoring function and required features count, the method returns features which are highly correlated with the class variable. The parameters *F classifier* is used for scoring function, *k* best is used for mode and *param* is set to 20. F classifier performs ANOVA F-value test between a feature and class variable. *Param* is the number of features which is returned. *K* best selects the top k best features.

9) Select Percentile(F classifier): Select Percentile method is used to identify the best features with in the selected percentile. The *percentile* value parameter is set to 10 and scoring function is F classifier. The method selects the top 10 percentile best features based on the F classifier statistical test.

10) Select K Best: This feature selection method identifies the k best features. The *k* value parameter is set to 15 and scoring function is F classifier. F classifier is used to perform statistical tests. This uses one way ANOVA F-test. The best

k values features are selected based on the statistical test.

11) Select False Positive Rate: Select False Positive rate is defined as a univariate filter-based feature selection method. It is the probability of falsely rejecting the null hypothesis. This statistical feature selection method is computed by ANOVA F-value statistic. The highest p values obtained from this test are selected. The *alpha* parameter is set to 0.01, which is the limit for the highest p-value for the features to be identified.

12) Select False Discovery Rate: It is a univariate statistical test for each of the features. FDR is defined as the error rate, which is the proportion of the false positives of the entire rejected hypothesis. Benjamini and Hochberg's paper introduced the ordering of p-value and proved that this ordering controls the FDR. Feature selection is performed using FDR in the study (Kim, Chen, Park, Ziegler, & Jones, 2008) and it is observed that on experimentation with human plasma the effectiveness of the method outperformed other statistical methods and showed better classification accuracy while keeping the relevant features. The *alpha* parameter is set to 0.01, which is the limit for the highest p-value for the features to be identified.

C. Classifier

In this study we have used three machine learning models namely Support Vector Machine(SVM), Neural Network (NN) and Random Forest(RF).

1) Neural Network: Artificial neural network is a computational model that replicates the parallel nature of neurons in a human brain. A layer in this network constitutes a subgroup of processing elements, where the initial layer is the input layer and the final layer constitutes the output. Between these layers, there are many hidden layers. Input layer neurons have a value that is to be propagated to the last layer through the hidden layers. The weights of neurons are updated at each iteration based on the back propagation method. In backpropagation, the value at the final layer is compared with the predicted values. Based on the mean squared error between the two values the error weight is propagated back to each of the hidden layers.

2) Support Vector Classification: Support vector machine (SVM) is highly popular in medical diagnosis as it provides an excellent performance in generalization (Novakovic & Veljovic, 2011). SVM plots the data instances in a higher dimensional space and constructs a hyperplane to classify the instances. The maximum margin hyperplane is chosen in a way that the distance from the hyperplane to the closest point on either side of the plane is maximized as shown in the figure. If the data in hyperplane is not linearly separable, the original dimensional space is mapped into higher dimensional space using a kernel function. Figure 2. shows the sample instances of two features in the feature space separated by a support vector such that the distance between the support vector and the to the closest point is maximized.

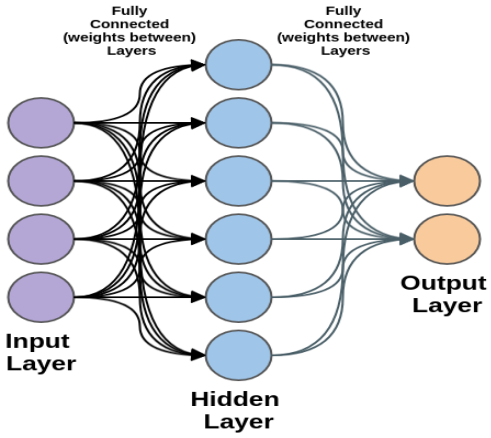

Figure 1. Neural Network

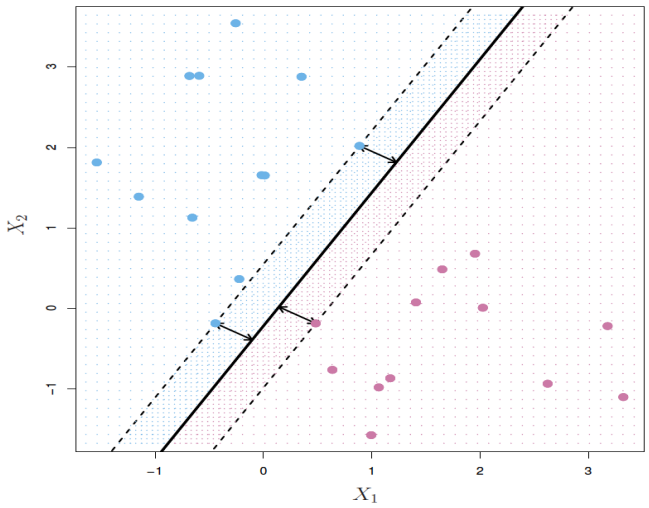

Figure 2. Support Vector Machine

3) Random Forest: Random Forests are ensemble model with constitutes of decision trees. These trees fit to bootstrap resamples of data and all the results from the trees are aggregated into one, this process is bagging. The accuracy in prediction is improved considerably due to tree combination algorithm where the most popular class is selected based on the input of the individual trees. This classifier can handle both continuous and discrete values. The other advantages of using this model are that it is very robust, fast, easy to use and accurate. Random forests are extensively used in bioinformatics and medicine. In the study(Parmar, Grossmann, Bussink, Lambin, & Aerts, 2015) machine learning methods for radiomics more than 12 machine learning classifiers are observed and concluded that random forest displayed the highest predictive performance. Majority of the feature selection methods in this study gave the highest performance when combined with random forest classifier.

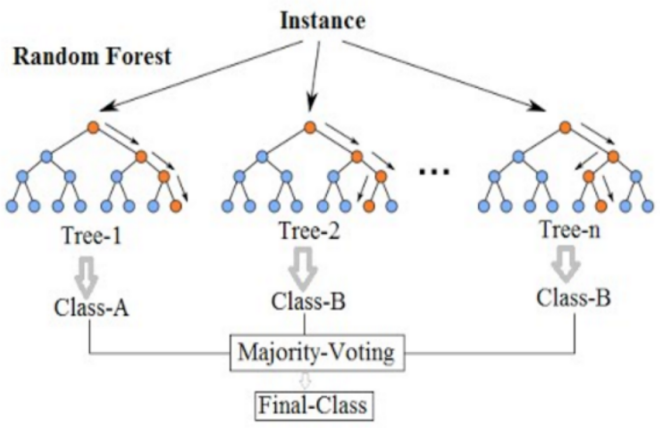

Figure 3. Random Forest

IV. Experimental setup

After preprocessing, four types of training dataset were identified namely non-imputed without smote, non-imputed with smote, weighted-knn without smote, and weighted-knn with smote. The experiment is carried out for 50 times for each identified datasets i.e., 50 fold validation with replacement (each time a different combination of instances were selected). For each training datasets following steps were carried out to identify top influential features for the classifiers namely Neural Nets, SVM, and Random Forest.

Step 1: For various datasets, in each iterations different combination of training instances are selected. All feature selection methods are run for 50 times to identify the top 15 features.

Step 2: In each iteration, feature selection methods returns scores for each identified features, there scores are cumulated. Each feature selection method has a different approach for scoring. Few methods rate a feature with a positive number, higher the number the more important is the feature, few methods rates a feature as 1 if it's important and 0 if it's not important.

Step 3: At the end of 50 iterations, the cumulative scores for each feature are averaged and ranked in ascending order. As the rank is based on the cumulative feature importance score, the most repeated feature with a higher score will have a higher rank. By this the top 15 features are extracted and returned for each feature selection method.

Step 4: The extracted feature sets from the previous step, are further filtered to identify the optimal feature number which has the best AUC and Accuracy for the classifiers. Feature combination is iteratively constructed from feature length 2 upto 15. The first combination consists of the top 2 ranked features from the list, while the second combination consists of the top 3 ranked features from the list. This combination generation is continued until the top 15 features. For each classifier the combinations are run for 50 fold times with repetition. The results from each iteration are averaged and the combination which gives highest result is extracted for each classifier.

Step 5: A graph is plotted for the feature selection method which gives the highest result for a particular classifier. The feature count is displayed in the x-axis and AUC on y-axis. Number of optimal features for a model are manually selected based on the graph. It is selected when the AUC is highest or there is no significant increase with an increase in feature length. To keep low dimensionality, priority is given for less feature length.

Step 6: A statistical test is applied for the highest performing results for each classifiers. This test is critical in machine learning to establish if there is significant difference between the different models. In this study Anova is used to validate the results.

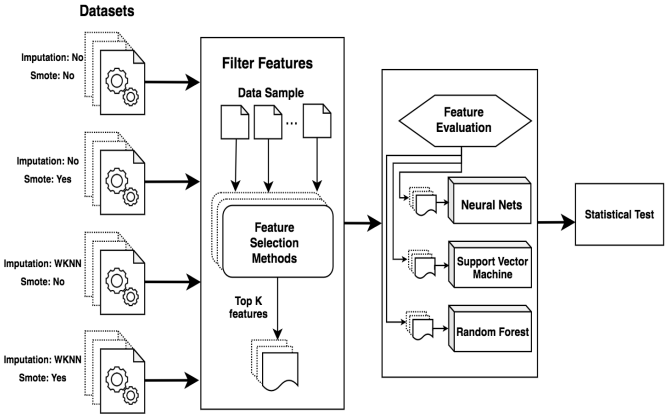

Figure 4. Overview of the experimental setup

V. Evaluation Metrics

Through the empirical study, it is difficult to establish the metric that is to be used for a problem as each of the evaluation methods have features to measure, the aspects of the model to be evaluated. Evaluating models using clinical data is difficult as there are large weighted discrepancies that arise from the difference between the actual value and predicted value(Aftarczuk, 2007). The evaluation of models using clinical data is a tradeoff between the true positive and true negative rate. The graphical representation of the tradeoff is through the Receiver Operating Characteristic(ROC) curve. On the X-axis of this curve is the percentage of False positive and the Y-axis represents the percentage of a true positive. The accepted performance metric for this ROC curve is the Area Under Curve (AUC). The model is evaluated based on the AUC, where higher the area under the curve represents a better model. This is a useful metric as it is independent of the prior probabilities and decision criteria selected. The other metric used for evaluating the models in this scenario is Accuracy. It is defined as the percentage between correctly classified true positive and true negative by the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

To compare the different machine learning models statistical test Anova is performed. Anova is an analysis of variance which determines if the mean difference between the models is zero. The null hypothesis in this test is that there is no significant difference between the model results, that is the mean difference is zero. While the alternate hypothesis is that there is a significant difference between the model results if the mean difference is not zero. The confidence level for Anova test is set to 95 percent. Conventional acceptance of a statistical significance is at a P-value of 0.05 or 5% and the class intervals are calculated at a confidence of 95% (Altman, Gore, Gardner, & Pocock, 1983). If the Anova determines that there is a statistically significant difference between the model results, then a post hoc test is carried out. Tukey's honestly significant difference(HSD) post hoc test is implemented to find which model results vary and how much the mean value differ. In Tukey test all possible combination of mean values were compared.

VI. Results

The real world dataset obtained from Doherty Institute had more than 10% of missing value. On consultation with subject experts case deletion method is used to delete the 15 attributes which had more than 60 percent missing data. All the instances which had a missing value is deleted resulting in 152 instances and 150 attributes. This is used as a original dataset (dataset 1). The missing values are filled using imputation methods of k-nearest neighbours for continuous data and weighted KNN with attribute association for categorical data. SMOTE is applied to remove the imbalance in number of classes in the original dataset. Four different datasets are created with the combination of choice of imputation and smote as given in the Table I.

Table I
Four different dataset used in the study

Dataset	Imputation	SMOTE
1	✗	✗
2	✗	✓
3	✓	✗
4	✓	✓

For each dataset, the selected twelve feature selection methods are applied to identify the top fifteen influential features. Some of the features like infant length at six months, infant weight at 6 months, mother's height, mother's weight, mother's education, infant's sex and wealth index has been identified by most feature selection methods as top fifteen features for a given dataset.

For each dataset and feature selection method, subset of features are incrementally selected ranging from 1 up to 15 from the corresponding top 15 features. These subset of features are then evaluated by three selected classifiers. Figure 5 shows three graphs plotted for each classifier between number of features in the subset and their corresponding AUC.

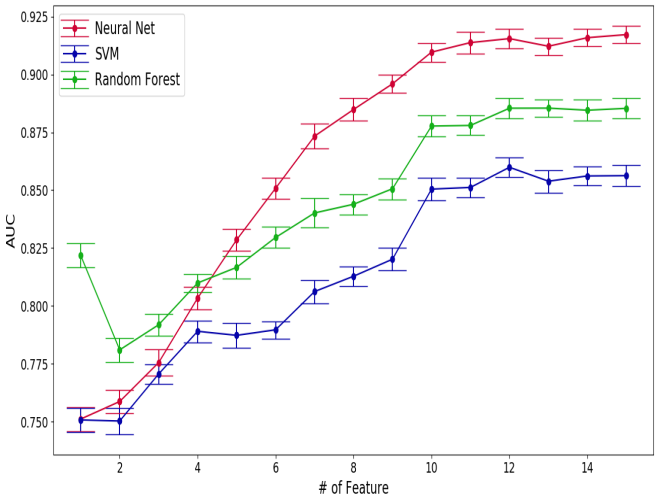

Figure 5. Mean AUC of classifiers over # of features for a dataset with imputation and smote applied where X-axis is # of features and Y-axis is mean AUC

It is observed that there is a significant increase in AUC when the number of features incremented from 1 to 10 and no significant increase in AUC after that. The subset of features to be used is identified based on the similar cutoff. For each dataset and the selected classifier, the best feature selection method is identified based on mean AUC and its standard deviation. The feature selection methods along with the mean AUC and standard deviation evaluated by Neural Nets on the dataset where imputation and SMOTE are applied is given in Table II.

From the Table II, it is observed that "Select False Positive Rate" method performs well with high AUC and less standard deviation than other feature selection methods. Similarly best feature selection methods are identified for each classifier on a given dataset. The list of such best feature selection method for each classifiers for a given dataset is given in Table III.

For each dataset, three models are created with the combination of three classifiers and their respective best performing feature selection method. The performance of the models are evaluated with cross validation of training dataset. The mean AUC and mean accuracy with standard deviation of 95% confidence interval is shown in Table III. The models are validated using a test dataset provided by Doherty Institute and the corresponding AUC and accuracy are provided in Table III. To determine if there is a statistically significant difference in models for a given dataset, ANOVA method is applied and the corresponding p-value is provided in Table III. The impact of imputation on the dataset is observed from the table. Even Though there is only a slight increase in AUC when the imputation is applied, the standard deviation of AUC has decreased by a large margin. This implies that the model performs consistently well when imputation is applied as there

Table II
PERFORMANCE OF FEATURE SELECTION METHODS USING NEURAL NETWORKS FOR A DATASET WHERE IMPUTATION AND SMOTE ARE APPLIED

Feature selection method	CI AUC	AUC stddev
Select False Positive Rate	0.92 ± 0.004	0.0121
Select False Discovery Rate	0.92 ± 0.004	0.0128
Mutual Information	0.92 ± 0.004	0.0129
Select K best	0.92 ± 0.004	0.0136
Generic Univariate Select	0.92 ± 0.004	0.0144
Select Percentile	0.92 ± 0.003	0.0147
Mutual Information Maximisation	0.92 ± 0.003	0.0126
Joint Mutual Information	0.92 ± 0.004	0.0142
Conditional Mutual Information Maximisation	0.92 ± 0.004	0.0138
Feature Importance	0.91 ± 0.003	0.0126
L1 based feature selection	0.88 ± 0.004	0.0147
Minimum Redundancy Maximum Relevance	0.86 ± 0.004	0.0159

are more instances from which the models can learn more and can predict stunting with high AUC consistently.

The effect of applying SMOTE on the dataset is also evident from the results shown in Table III. SMOTE creates a boundary with the class instances that are less in number and creates more such instances within the boundary. This is done to remove the imbalance in number of classes in training instances. This prevents the model from overfitting and improves the effectiveness in predicting both classes with accuracy. This is evident from the results shown in Table III, where there is a significant increase in AUC when SMOTE is applied.

The models are validated with the test dataset which contained 163 instances with 12 instances of stunting36 class as yes and rest of the instances are classified as no. The performance of the models has decreased significantly on this test data. This decrease in performance is due to the imbalance in number of classes in the training dataset. The models, when applied on the training dataset with imbalance in class variables, overfit the data and tend to classify the class variable as no than yes.

It is observed that the issue of lower performance is not countered when SMOTE is applied to rebalance the dataset. The reason for this poor performance is that the test instances classified as yes fall outside the boundary created by SMOTE. The test instances which lies inside the SMOTE boundary are classified correctly. This is validated with the training dataset where only imputation is applied. This dataset contains 1168 instances out of which around 150 instances have been classified as yes. The no classified instances are randomly picked to match the instances which are classified as yes. When model is created on this dataset and evaluated with test data the AUC is 0.77. This is significantly higher when compared to dataset where the imbalance in class variables is present or when SMOTE technique is applied. The performance of the models in test data with this dataset is also similar to the performance of model in training dataset.

The three classifiers which is applied on a dataset is evaluated against each other using ANOVA. The null hypothesis in this test is that there is no significant difference between the classifiers. From the p-value given in the Table III, it is observed that the alternate hypothesis is true. That is the three classifiers are statistically significantly different from each other. The ANOVA test is followed by the post-hoc Tukey test whose results showed that neural networks perform significantly better than other two classifiers. In Figure 6 graphs are plotted for each classifier between probability distribution function and AUC. It is observed from graph that neural networks perform well with high mean AUC and less standard deviation than other two classifiers.

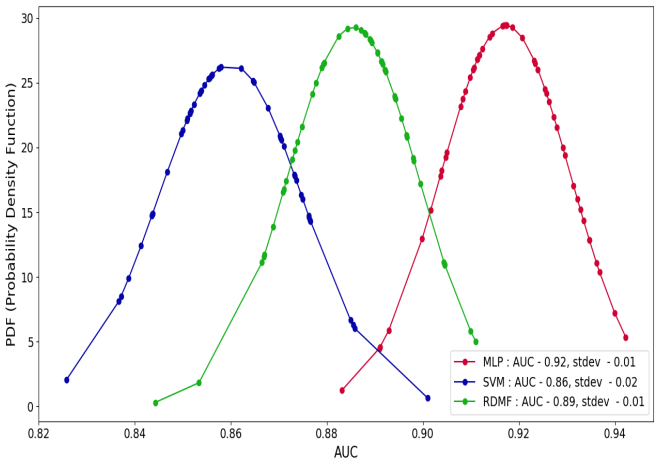

Figure 6. Mean population range of the classifiers over AUC range where X axis contains AUC and Y axis is the probability distribution function

VII. CONCLUSION

Stunting is a common health problem that is faced by children in Vietnam. In this study, we have applied machine learning methods and feature selection techniques to identify the determinants. The dataset provided by the Doherty institute had many missing values. These missing values had to be handled for effective analysis. The distribution of classes in the dataset is imbalanced. To handle the imbalance SMOTE is implemented on the dataset and synthetic values are imputed to increase the stunting class variables. The dataset has continuous and categorical data, knn imputation is performed on the continuous data and weighted knn on the categorical data. On this imputed dataset one-hot, encoding and scalar standardisation is performed to prepare it for machine learning methods. The datasets with implemented weighted knn and SMOTE had better performance compared to the other datasets. The standard deviation while using this dataset was minimum compared to the other three datasets.

Table III
EVALUATION RESULTS - MEAN AUC AND ACCURACY WITHIN 95% CI FOR MODELS WITH BEST FEATURE SELECTION METHOD AND CLASSIFIER FOR A GIVEN DATASET. THE P-VALUES TO COMPARE STATISTICAL DIFFERENCE BETWEEN CLASSIFIERS FOR GIVEN DATASET

Imputation	Smote	Classifier	Feature selection method	AUC	Accuracy	Test AUC	Test Accuracy	P-value
None	No	Neural Nets	Select False Discovery Rate	0.78 ± 0.033	0.87 ± 0.033	0.58	0.85	1.12E-11
None	No	Support Vector Machine	Select Percentile	0.64 ± 0.026	0.84 ± 0.026	0.54	0.85	
None	No	Random Forest	L1 based	0.63 ± 0.028	0.81 ± 0.028	0.59	0.86	
None	Yes	Neural Nets	Select False Positive Rate	0.95 ± 0.009	0.95 ± 0.009	0.58	0.85	8.10E-09
None	Yes	Support Vector Machine	Select False Discovery Rate	0.92 ± 0.011	0.92 ± 0.011	0.60	0.83	
None	Yes	Random Forest	Generic Univariate Select	0.90 ± 0.012	0.90 ± 0.012	0.62	0.87	
Wknn	No	Neural Nets	Select Percentile	0.66 ± 0.012	0.85 ± 0.012	0.51	0.83	4.98E-18
Wknn	No	Support Vector Machine	Select False Positive Rate	0.59 ± 0.008	0.87 ± 0.008	0.53	0.86	
Wknn	No	Random Forest	Select False Positive Rate	0.62 ± 0.009	0.86 ± 0.009	0.51	0.85	
Wknn	Yes	Neural Nets	Select False Positive Rate	0.92 ± 0.004	0.92 ± 0.004	0.55	0.84	5.01E-11
Wknn	Yes	Support Vector Machine	Feature Importance	0.90 ± 0.003	0.90 ± 0.003	0.65	0.84	
Wknn	Yes	Random Forest	Feature Importance	0.91 ± 0.004	0.91 ± 0.004	0.57	0.86	

Feature selection methods have been used for high-throughput data mining problem. In this study, we have used 12 feature selection methods which predict stunting. Filter methods are used in this study as they are independent, computationally efficient and are prone to less overfitting than embedded or wrapper feature selection methods. Our results show that univariate feature selection performed best with all classifiers. The best feature selection method identified was SelectFPR. The features identified by this method are: infant length at 6 months, infant weight at 6 months, mothers weight throughout the pregnancy period, mothers height, infant's weight at birth, father's weight, father's height, mothers education, wealth index of the family and mothers job. These features align with the features identified in literature reviews and the features identified by Doherty Institute through regression. These determinants can be checked on from early stages and can reduce the possibility of stunting in a child. Machine learning models are compared to each other based on statistical ANOVA test. From this test, it is concluded that there is a significant difference between classifiers and neural networks performs the best.

In our future works, we will incorporate deep learning into the study, further more sophisticated imputation methods can be implemented. Our study identifies the determinant features for stunting by using univariate and multivariate feature selection methods. Various embedded and wrapper methods can be implemented to identify and compare the best features which classify stunting.

REFERENCES

Aftarczuk, K. (2007). Evaluation of selected data mining algorithms implemented in medical decision support systems..

Ahmed, A. S., Ahmed, T., Roy, S., Alam, N., & Hossain, M. I. (2012). Determinants of undernutrition in children under 2 years of age from rural bangladesh. *Indian Pediatrics*, *49*(10), 821?824. doi: 10.1007/s13312-012-0187-2

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using smote and ensemble machine learning approach: The henry ford exercise testing (fit) project. *Plos One*, *12*(7). doi: 10.1371/journal.pone.0179805

Ali, D., Saha, K. K., Nguyen, P. H., Diressie, M. T., Ruel, M. T., Menon, P., & Rawat, R. (2013). Household food insecurity is associated with higher child undernutrition in bangladesh, ethiopia, and vietnam, but the effect is not mediated by child dietary diversity. *The Journal of Nutrition*, *143*(12), 2015?2021. doi: 10.3945/jn.113.175182

Altman, D. G., Gore, S. M., Gardner, M. J., & Pocock, S. J. (1983). Statistical guidelines for contributors to medical journals. *Bmj*, *287*(6385), 132?132. doi: 10.1136/bmj.287.6385.132-a

Alzubi, R., Ramzan, N., Alzoubi, H., & Amira, A. (2018). A hybrid feature selection method for complex diseases snps. *IEEE Access*, *6*, 1292-1301. doi: 10.1109/ACCESS.2017.2778268

Battiti, R. (1994, July). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, *5*(4), 537-550. doi: 10.1109/72.298224

Blagus, R., & Lusa, L. (2015). Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. *BMC Bioinformatics*, *16*(1). doi: 10.1186/s12859-015-0784

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A. C., Grisel, O., ... et al. (2013, Sep). *Api design for machine learning software: experiences from the scikit-learn project.* Retrieved from https://arxiv.org/abs/1309.0238

Cerda, P., Varoquaux, G., & Kgl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, *107*(8-10), 1477?1494. doi: 10.1007/s10994-018-5724-2

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience.

Cruz, L. G., Azpeitia, G. G., Sarez, D. R., Rodrguez, A. S., Ferrer, J. L., & Serra-Majem, L. (2017). Factors associated with stunting among children aged 0 to 59 months from the central region of mozambique. *Nutrients*, *9*(5), 491. doi: 10.3390/nu9050491

Faisal, S., & Tutz, G. (2017, Oct). *Nearest neighbor imputation for categorical data by ...* Retrieved from https://arxiv.org/pdf/1710.01011.pdf

Geurts, P., Ernst, D., & Wehenkel, L. (2006, Apr 01). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. Retrieved from https://doi.org/10.1007/s10994-006-6226-1 doi: 10.1007/s10994-006-6226-1

Guyon, I., & Elisseeff, A. (2003, March). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*, 1157–1182. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944968

Hall, B. (2016). Facies classification using machine learning. *The Leading Edge*, *35*(10), 906?909. doi: 10.1190/tle35100906.1

Khare, S., Kavyashree, S., Gupta, D., & Jyotishi, A. (2017). Investigation of nutritional status of children based on machine learning techniques using indian demographic and health survey data. *Procedia Computer Science*, *115*, 338?349. doi: 10.1016/j.procs.2017.09.087

Kim, S. B., Chen, V. C. P., Park, Y., Ziegler, T. R., & Jones, D. P. (2008). Controlling the false discovery rate for feature selection in high-resolution nmr spectra. *Statistical Analysis and Data Mining*, *1*(2), 57?66. doi: 10.1002/sam.10005

Lartey, A. (2015). What would it take to prevent stunted growth in children in sub-saharan africa? *Proceedings of the Nutrition Society*, *74*(4), 449?453. doi: 10.1017/s0029665115001688

Lewis, D. D. (1992). Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on speech and natural language* (pp. 212–217). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from https://doi.org/10.3115/1075527.1075574 doi: 10.3115/1075527.1075574

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017, December). Feature selection: A data perspective. *ACM Comput. Surv.*, *50*(6), 94:1–94:45. Retrieved from http://doi.acm.org/10.1145/3136625 doi: 10.1145/3136625

Little, R. J. A., & Rubin, D. B. (2010). *Statistical analysis with missing data*. John Wiley & Sons, Inc.

Liu, S., & Motani, M. (2018). Feature selection based on unique relevant information for health data. *CoRR*, *abs/1812.00415*. Retrieved from http://arxiv.org/abs/1812.00415

Markos, Z. (2014). Predicting under nutrition status of under-five children using data mining techniques: The case of 2011 ethiopian demographic and health survey. *Journal of Health & Medical Informatics*, *5*(2). doi: 10.4172/2157-7420.1000152

Menon, P., Bamezai, A., Subandoro, A., Ayoya, M. A., & Aguayo, V. M. (2015). Age-appropriate infant and young child feeding practices are associated with child nutrition in india: insights from nationally representative data. *Maternal & Child Nutrition*, *11*(1), 73?87. doi: 10.1111/mcn.12036

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009, Jul 10). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, *10*(1), 213. Retrieved from https://doi.org/10.1186/1471-2105-10-213 doi: 10.1186/1471-2105-10-213

Novakovic, J., & Veljovic, A. (2011, Sep.). C-support vector classification: Selection of kernel and parameters in medical diagnosis. In *2011 ieee 9th international symposium on intelligent systems and informatics* (p. 465-470). doi: 10.1109/SISY.2011.6034373

Parmar, C., Grossmann, P., Bussink, J., Lambin, P., & Aerts, H. J. W. L. (2015). Machine learning methods for quantitative radiomic biomarkers. *Scientific Reports*, *5*(1). doi: 10.1038/srep13087

Prendergast, A. J., & Humphrey, J. H. (2014). The stunting syndrome in developing countries. *Paediatrics and International Child Health*, *34*(4), 250?265. doi: 10.1179/2046905514y.0000000158

Rubin, D. B. (2011). *Multiple imputation for nonresponse in surveys*. John Wiley.

Sarma, H., Khan, J. R., Asaduzzaman, M., Uddin, F., Tarannum, S., Hasan, M. M., ... Ahmed, T. (2017). Factors influencing the prevalence of stunting among children aged below five years in bangladesh. *Food and Nutrition Bulletin*, *38*(3), 291?301. doi: 10.1177/0379572117710103

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, *17*(6), 520?525. doi: 10.1093/bioinformatics/17.6.520

Zeng, X.-Q., & Li, G.-Z. (2014). Supervised redundant feature detection for tumor classification. *BMC Medical Genomics*, *7*(Suppl 2). doi: 10.1186/1755-8794-7-s2-s5

Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., & Keane, J. (2009). Comparing data mining methods with logistic regression in childhood obesity prediction. *Information Systems Frontiers*, *11*(4), 449?460. doi: 10.1007/s10796-009-9157-0