Forensic Prediction using Bias-Variance Tradeoff and Random Forest Algorithms

S Shanthini, S Vinu

Abstract: ***Every individual host after death has its own altered micro biome configuration. After death, postmortem microorganism communities change to represent the attributes of death. The micro biome act as a many roles in human health, usually done by the exclusive lens of clinical interest. By scouring 5 anatomical areas throughout regular demise exploration from 188 case to predict the Postmortem Interval (PMI), location of death and manner of death, the postmortem micro biomes were collected. The micro biome sequencing are not easy to analyze and interpret because it produces large multidimensional dataset. To overcome the analytical challenge Machine learning method can be used. The two supervised machine learning methods employed here are Random Forest and Bias-Variance Tradeoff. In training datasets, Random forest algorithm is applied. This algorithm makes predictions by choosing the most voted node from each decision tree as the output. The output is checked for any bias variance error, by the Bias-Variance Tradeoff algorithm in order to help the supervised learning algorithm to perform generalization beyond the training datasets. To obtain a prediction that is best fitted and accurate, these two algorithms are chosen for learning.***

Keywords: Bias-Variance Tradeoff, location of death, manner of death, Postmortem Interval, Postmortem microbiome, Random forest.

I. INTRODUCTION

In Research or educational purposes, to work out the cause, mode and manner of death or to judge any malady or injury,thorough examination of a dead body by dissection is needed . The word autopsy is derived from the Greek word ‘autopsia cadaverum’, which means “to see with one’s own eyes”. Network of microorganisms are presented in the human body. The micro biome is the genetic material of all microorganisms such as germs, fungus, protozoa and virus that occupy on and inside the individual body.To obtain valuable information on death investigation, the composition and individual microorganism have been broadly considered estimating individual physical condition. The host atmosphere, existence, or nonexistence of illness, growth and lifestyle markers are influenced by human micro biome. In a dynamic ecosystem, it will be changeable within an entity, as different consortium resides on or in different parts of the body. However, microbial biodiversity dynamics in human population is found to be consistent until 48 hours after death [3], after which they change in composition. In industrial-urban population cross-section [3] 5 anatomical areas are swabbed from 188 cases to get the information about varying microbial communities among anatomical areas and changes taking place after the death. The valuable information related condition of death can be found from rectum and eye since it takes longer amount of time to affect during decomposition. These micro biomes are the building blocks of molecular autopsy which is used as a diagnostic tool to identify gene mutation that determines the cause of death in unexplained cases.Using machine learning algorithms to forecast place of death, approach, moment and links with medical conditions based on the training datasets (microbial signatures). To understand the error in prediction is the big challenge in forensic investigation. The data about the behaviour of these microbial is multidimensional. However, these algorithms have certain complexities because they are “black boxes” wherein data is sent as input and an output is computed without the knowledge of the inner workings of the system. In order to understand the predictions and eliminate the errors effectively , two machine learning methods- Random forests and Bias-Variance Tradeoff are used.

II. METHODOLOGY

The microbial behavior observed within and after 48 hours of death for 188 death cases [1] are given as training data into the machine learning model. Data with low variance is processed by using Random forest method. Numerous decision trees are generated by the algorithm. In this model,

- For every tree, classes or mean prediction are taken.
- All these predictions from individual trees obtained from individual datasets are sent as input into the bias variance model to be corrected of bias and variance error.
- The final outcome is found to be closest to accurate, as the Random forest algorithm generates the most voted daughter node as the final output, even though the dataset is multidimensional.

III.RANDOM FOREST ALGORITHM

In machine learning, most popular and powerful supervised algorithm to perform both regression and classification tasks is Random Forest. In Random forest,

- Forest can be created with number of decision trees.
- The prediction will be robust when there are more no of trees within the forest.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author

S Shanthini*, Assistant Professor, CSE Department, St.Joseph’s College of Engineering, Chennai, India.

S Vinu, Assistant Professor, CSE Department, St.Joseph’s College of Engineering, Chennai , India.

- ❖ Accuracy of the result is high in random forest classifier when more no of tree in the forests.
- ❖ Using Information gain and Gigi index algorithm, multiple decision trees are constructed.
- ❖ For Classification, attributes are selected for new object and based on majority of the votes for the trees; classes can be found [5].
- ❖ The final class in the tree is chosen from the majority voting of all the other trees. The overview of random forest algorithm is shown in Fig.1.

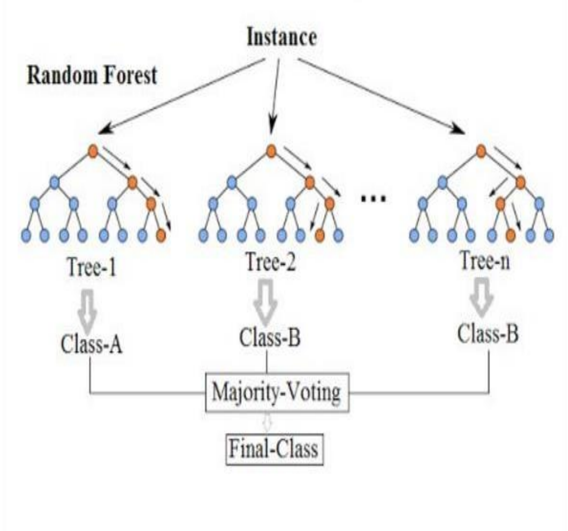


Fig.1.Random Forest Simplified

Ensemble Machine Learning algorithm based on divide and conquer approach used to improve performance. In ensemble method, Strong learners can be formed from group of weak learners. Thus, to reduce the variance and improve performance in method is used.

IV. BIAS VARIANCE TRADEOFF

In learning algorithms, according to erroneous predictions the Bias error is an error. High bias will result in not getting correct significant associations among features and intented output (underfitting). Variance quantifies the difference in predictions when we change the dataset. If we have high variance, it means that one prediction is going to be different when we give the same test case. Two sources of error are minimized simultaneously by the contradictory Bias Variance problem in order to prevent learning algorithm from generalize outside the training set.The fig. 2 depicts the relationship between bias, variance and the accuracy of the final prediction [2]. Three important terms variance, bias and a quantity called the irreducible error are used to reduce the noise in the problem. Expected generalization error with respect to a particular problem can be find using Bias Variance decomposition .Select a good learning algorithm to reduce the bias and variance value. Most important properties of estimators are Bias and in the training dataset, variance is defined as an error from sensitivity to small fluctuation. So it will create random noise in the training data. Hence, Bias-Variance Tradeoff is necessary.

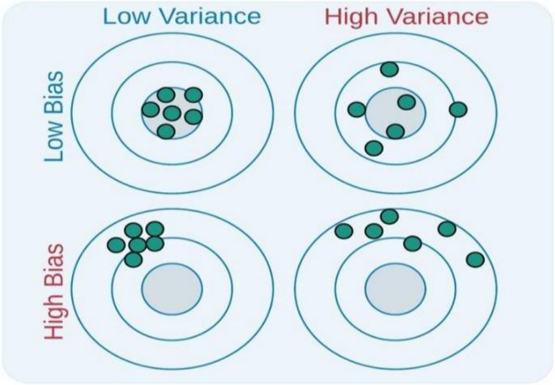


Fig.2.The relationship between bias and variance

V.PREDICTION

Random forest algorithm is used to find the prediction and the sample prediction with five anatomical areas from three combinations of predictor variables are given in the table I, II and III below [1].

Table-I: Prediction of Postmortem Interval

Predictor Variable	Prediction/ Observation	<24 h	25-48 h	49-72 h	>73 h
Postmortem Interval Estimate	<24 h	300	63	28	06
	25-48 h	76	288	18	26
	49-72 h	03	02	20	01
	>73 h	01	00	00	15

Table-II: Prediction of Event Location

Predictor Variable	Prediction/ Observation	Hospital	Indoors	Outdoors	Vehicle
Event Location	Hospital	65	00	00	00
	Indoors	42	586	39	19
	Outdoors	05	03	67	06
	Vehicle	00	00	00	13

Table –III: Prediction of Manner of Death

Predictor Variable	Prediction/ Observation	Accident	Homicide	Natural	Suicide
Manner of Death	Accident	296	11	18	38
	Homicide	03	141	03	00
	Natural	37	09	221	26
	Suicide	01	01	03	39

VI. PREDICTOR VARIABLE AND PERFORMANCE METRIC

Accurateness and p-value are obtained from 5-fold cross validation i.e, rectum, ear, mouth , nose and eyes for Random Forest algorithm [1]to guess the location, manner of death and postmortem interval, are given in Table IV.

Table- IV: Accuracy and p-value [1]

Predictor Variable	Performance Metric	Random Forest
Postmortem Interval Estimate	Accuracy	0.736
	p-value	<2.0e-16
Event Location	Accuracy	0.863
	p-value	<2.0e-16
Manner of Death	Accuracy	0.823
	p-value	<2.0e-16

In the tables described above, we observe that the result is not at its best predictive accuracy. So, we apply Bias-Variance Tradeoff algorithm to remove bias and variance error. On the whole, applying Random forest algorithm and performing error detection using Bias-Variance Tradeoff improves the overall accuracy of the prediction.

VII. RESULT AND DISCUSSION

In this paper, we have taken prediction with three combination by swabs event place, postmortem interval, and nature of death. The input that provide is in the form of CSV file on implementation for a given number of samples from five subareas (anatomic areas) and get the most accurate model (highest accuracy) as outputs in Fig 3, 4 and 5. Accuracy depended on the technique applied, the numbers of anatomical areas analyzed, and the predicted characteristic of death.

A. Postmortem Interval Prediction Graph

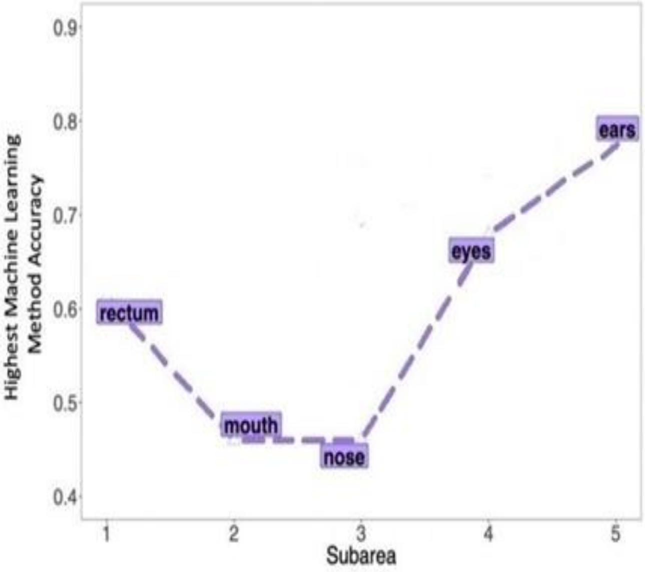


Fig 3.Postmortem Interval Prediction

B.Event Location Prediction Graph

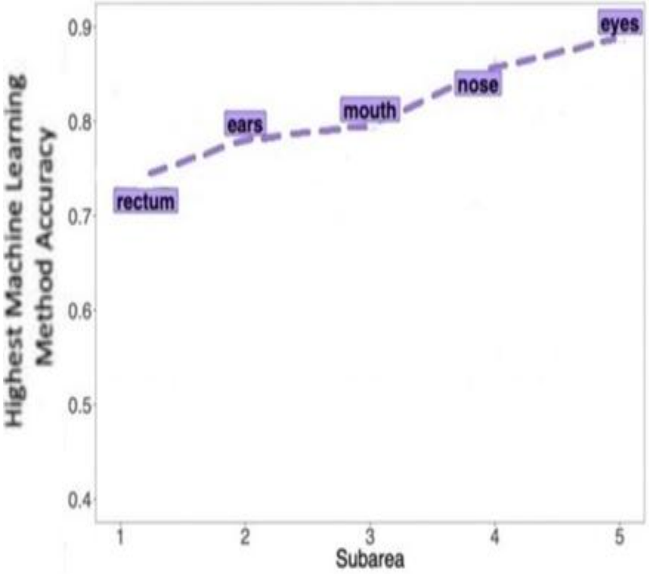


Fig 4.Event Location Prediction

C.Manner of Death Prediction Graph

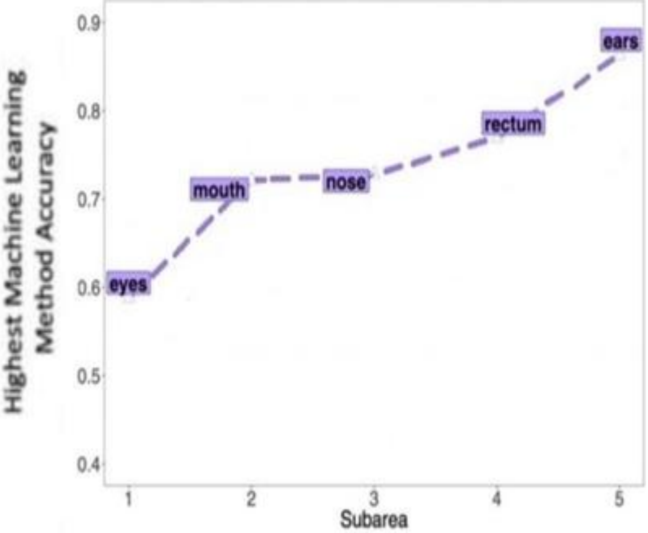
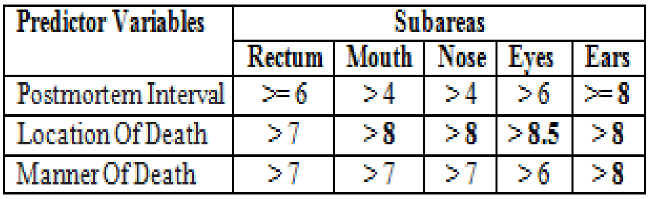


Fig 5. Manner of Death Prediction



Predictor Variables	Subareas				
	Rectum	Mouth	Nose	Eyes	Ears
Postmortem Interval	>=6	>4	>4	>6	>=8
Location Of Death	>7	>8	>8	>8.5	>8
Manner Of Death	>7	>7	>7	>6	>8

Random forest is the perfect algorithm to implement in this system. The accuracy achieved by Random forest was found to be 74.5%-87.6% by choosing three important attributes during death exploration to predict using the postmortem microbiota. Eyes were the most informative in predicting location of death, Ears in Postmortem Interval and manner of death.

VIII. CONCLUSION

Using Random forest, this model is able to handle large datasets with higher dimensionality. The problems encountered while implementing other supervised learning algorithms like Xgboost and neural networks have been overcome by this model. Random forest performed better with more numbers of anatomical areas in the model with higher sensitivity for less frequent classes. Eyes and Ears giving valuable information related with the investigation of death. Random forest handles the missing values and maintains accuracy for missing data. Overfitting and underfitting are observed to ensure perfect fit.

REFERENCES

1. Zhang Y, Pechal JL, Schmidt CJ, Jordan HR, Wang WW, Benbow ME, Machine learning performance in a microbial molecular autopsy context: A cross-sectional postmortem human population study. et al. (2019) PLoS ONE 14(4)
2. V. Gudivada, A. Apon, and J. Ding. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations". In: International Journal on Advances in Software 10.1 (2017), pp. 1 - 20.
3. Jennifer L. Pechal , Carl J. Schmidt, Heather R. Jordan ,M. Eric Benbow , A large-scale survey of the postmortem human microbiome, and its potential to provide insight into the living health condition Scientific Reports, (2018); pp.8: 5724
4. Aeriel Belk, Zhenjiang Zech Xu, David O. Carter, Aaron Lynne, Sibyl Bucheli, Rob Knight and Jessica L. Metcalf1, Microbiome Data Accurately Predicts the Postmortem Interval Using Random Forest Regression Models, Genes(2018),pp. 9
5. Subramaniam, Prashanth and Maninder Jeet Kaur. “Review of Security in Mobile Edge Computing with Deep Learning.” 2019 Advances in Science and Engineering Technology International Conferences (ASET) (2019), pp.1-5.
6. http://www.deathreference.com/Py-Se/Rigor-Mortis-and-Other-Postmortem-Changes.html

AUTHORS PROFILE

S.Shanthini got her B.E. degree in Computer Science & Engineering from Cape Institute of Technology and M.E degree in Computer Science & Engineering from SSN College of Engineering. Now, working as an Assistant Professor in St.Joseph’s College of Engineering, Chennai. She has more than 8 years experience in teaching. She interested in Networks, Database Management Systems, Data Structures, Algorithms and Data Mining.

S Vinu got his B.E. degree in Computer Science & Engineering from St.Xavier’s College of Engineering and M.E degree in CSE from St.Joseph’s College of Engineering. Now, working as an Assistant Professor in St.Joseph’s College of Engineering, Chennai. Currently doing his researches in Distributed systems and other areas includes Operating Systems, Algorithm, Python. He has more than 6 years experience in teaching.