Prediction of West Nile Virus using Ensemble Classifiers

R. Rishickesh, A. Shahina, A. Nayeemulla Khan

Abstract*: *West Nile Virus (WNV) is a disease caused by mosquitoes where human beings get infected by the mosquito's bite. The disease is considered to be a serious threat to the society especially in the United States where it is frequently found in localities having water bodies. The traditional approach is to collect the traps of mosquitoes from a locality and check whether they are infected with virus. If there is a virus found then that locality is sprayed with pesticides. But this process is very time consuming and requires a lot of financial support. Machine learning methods can provide an efficient approach to predict the presence of virus in a locality using data related to the location and weather. This paper uses the dataset present in Kaggle which includes information related to the traps found in the locality and also about the information related to the locality's weather. The dataset is found to be imbalanced hence Synthetic Minority Over sampling Technique (SMOTE), an upsampling method, is used to sample the dataset to balance it. Ensemble learning classifiers like random forest, gradient boosting and Extreme Gradient Boosting (XGB). The performance of ensemble classifiers is compared with the performance of the best supervised learning algorithm, SVM. Among the models, XGB gave the highest F-1 score of 92.93 by performing marginally better than random forest (92.78) and also SVM (91.16).

Index Terms*: *West Nile Virus, Ensemble Learning Classifiers, Extreme Gradient Boosting, Synthetic Minority Over sampling Technique.

I. INTRODUCTION

West Nile Virus (WNV) is a mosquito-borne disease that infects a human being through an infected mosquito's bite. The disease is commonly found in the United States and in recent times, has been reported in some places in India. It occurs frequently during the mosquito season that begins in the summer and proceeds till the fall. Very less affected people show symptoms of the infection i.e. only as few as 20% of the total population affected get viral fevers and other serious illness. Only 0.6% of the people affected get fatal illness. Still, the epidemic is considered to be a serious threat to the society and preventive measures are taken to control its spread. So far there has been no specific treatment like vaccinations and medicines for the WNV. But prevention measures like wearing long sleeve shirts and long pants are taken to reduce risk of getting mosquito bites.

A city has many localities, especially the ones which have a river, that are affected by this seasonal epidemic. These localities must be protected in advance by spraying pesticides in the form of smoke to control the adult mosquito populations. The smoke consisting of the pesticide is very harmful even to human beings. The mosquitoes in every locality are tested for the presence of the virus in the mosquitoes and if it is positive then the area to which the trap of mosquitoes belong are sprayed with pesticides. This is a very tedious approach which requires a lot of time and huge money invested. Therefore machine learning approaches can be used to predict the presence of the virus in an economical way.

Mitch Campion et.al, predicted the trap count of the species *Culex Tarsalis* [1], which frequently causes the WNV in North Dakota, using machine learning methods. The data used in that work was collected from the meteorological department of North Dakota which includes the trap count data of *Culex Tarsalis* from 2005-2015. The geographical data included rainfall, temperature, relative humidity, etc. The technique used in that work to predict the trap count was partial least square regression. The Mean Absolute Error (MAE) value was used as the statistical comparison measure to analyze the efficiency of the model. The dataset used in this work does not contain physical factors like average dew point, average pressure and the time when sunrise or sunset takes place along with the location features like the address and the latitude-longitude coordinates. These features give information about the number of mosquitoes present in each trap.

Eliza Little et.al, proposed an ensemble model to predict the WNV infection rates in the *Culex* species among the species present in Suffolk County, New York [2]. The data used is collected from the meteorological department which also includes hydrological data. The model is trained using the data from the year 2001 to 2009, and then a set of models are built using the data from 2001-2012 to validate the model. Using the already built ensemble model the prediction for the year 2013-2015 is done. The 12th model (m12) suggested that the warmer conditions are the reason behind increasing the infection rate in the *Culex* mosquitoes.

In earlier works they have predicted the trap count of the species causing WNV virus [1] using very few physical factors and the infection rate in the *Culex* species [2]. In the present work physical factors like average dew point, average pressure and the time when sunrise or sunset takes place along with the location features like the address and the latitude-longitude coordinates are used to predict the presence of WNV in different localities since they give good information about the number of mosquitoes present in each trap. By having the weather, location and trap count data, we can predict whether the epidemic will rise or not in a particular area using machine learning and data mining technique [3].

Revised Manuscript Received on October 15, 2019.

R. Rishickesh, Department of Information Technology, SSN College of Engineering, Kalavakkam-603110, India

A. Shahina, Department of Information Technology, SSN College of Engineering, Kalavakkam-603110, India .

A. Nayeemulla Khan, School of Computing Science and Engineering, VIT University, Chennai-600127, India

This approach consumes very less time and is less tedious than the traditional methods. In this work, ensemble algorithms like random forest, gradient boosting and eXtreme Gradient Boosting (XGB) in addition to Support Vector Machine (SVM) are considered for modeling. The performance metric used in this approach is the F-1 score [4]. The F-1 score of each algorithm is compared and the best model is decided.

II. DATASET

The dataset used in the present study is the "West Nile Virus Prediction" is given by Kaggle. The dataset was officially released by the Chicago Health department in the year 2015.

III. EXPLORATORY DATA ANALYSIS

The data repository contains 2 main datasets- *train.csv* and *weather.csv*. It also contains 2 other datasets which are inconsequential to our present work. The *train.csv* consists of the following features: The date on when the WNV test was performed, the address which gives the location of the traps collected for testing, species of the mosquitoes, the block number and street name present in the address. Along with them the other features used are, the unique identification number of the trap, the approximate address which is returned by the GeoCoder after passing the original address to it, the latitude, longitude, address accuracy values returned by the GeoCoder, the number of mosquitoes found in the trap and a binary value indicating whether WNV is present in the locality or not. The data collected in *train.csv* was from the year 2007-2014. The number of rows present in the dataset is 10506. The *weather.csv* dataset consists of all the weather factors from 2007-2014. It includes the maximum temperature, minimum temperature, average dew point, average pressure, etc.

	Date	Address	Species	Block	Street	Trap	AddressNumberAndStreet	Latitude	Longitude	AddressAccuracy	NumMosquitos	WnvPresent
0	2007-05-29	4100 North Oak Park Avenue, Chicago, IL 60634, ...	CULEX PIPIENS/RESTUANS	41	N OAK PARK AVE	T002	4100 N OAK PARK AVE, Chicago, IL	41.954690	-87.800991	9	1	0
1	2007-05-29	4100 North Oak Park Avenue, Chicago, IL 60634, ...	CULEX RESTUANS	41	N OAK PARK AVE	T002	4100 N OAK PARK AVE, Chicago, IL	41.954690	-87.800991	9	1	0
2	2007-05-29	6200 North Mandell Avenue, Chicago, IL 60646, USA	CULEX RESTUANS	62	N MANDELL AVE	T007	6200 N MANDELL AVE, Chicago, IL	41.994991	-87.769279	9	1	0
3	2007-05-29	7900 West Foster Avenue, Chicago, IL 60656, USA	CULEX PIPIENS/RESTUANS	79	W FOSTER AVE	T015	7900 W FOSTER AVE, Chicago, IL	41.974089	-87.824812	8	1	0
4	2007-05-29	7900 West Foster Avenue, Chicago, IL 60656, USA	CULEX RESTUANS	79	W FOSTER AVE	T015	7900 W FOSTER AVE, Chicago, IL	41.974089	-87.824812	8	4	0

Fig.1 First 5 rows of the train.csv dataset.

	Station	Date	Tmax	Tmin	Tavg	Depart	DewPoint	WetBulb	Heat	Cool	Sunrise	Sunset
0	1	2007-05-01	83	50	67	14	51	56	0	2	0448	1849
1	2	2007-05-01	84	52	68	M	51	57	0	3	-	-
2	1	2007-05-02	59	42	51	-3	42	47	14	0	0447	1850
3	2	2007-05-02	60	43	52	M	42	47	13	0	-	-
4	1	2007-05-03	66	46	56	2	40	48	9	0	0446	1851

Fig.2 The first 5 rows of the weather.csv dataset.

	Species	Block	Street	Trap	Latitude	Longitude	AddressAccuracy	month	day	Lat_int	Long_int	Tmax_x	Tmin_x	Tavg_x	Depart_x	DewPoint_x
0	2	41	32	1	41.954690	-87.800991	9	5	29	41	-87	88	60	74	10	58
1	3	41	32	1	41.954690	-87.800991	9	5	29	41	-87	88	60	74	10	58
2	3	62	27	6	41.994991	-87.769279	9	5	29	41	-87	88	60	74	10	58
3	2	79	109	13	41.974089	-87.824812	8	5	29	41	-87	88	60	74	10	58
4	3	79	109	13	41.974089	-87.824812	8	5	29	41	-87	88	60	74	10	58

Fig.3 Represents the dataset which is the resultant of merging the *train.csv* and *weather.csv* (joined on 'Date' column). The figure shows the first 16 columns of the dataset.

The *train.csv* contains information related to each area and number of mosquitoes present in that area on a particular date as shown in Fig.1. The column *WNVPresent* denotes whether the virus is present in that area or not. The *weather.csv* consists of the weather conditions of the locality on a particular date as shown in Fig.2. Both the datasets are joined to create a new dataset which is later used for modeling as shown in Fig.3.

The columns *Species*, *Trap* and *Street* are encoded with numbers since they are in the form of string and are categorical. So to insert them into a model they need to be assigned or categorized using numbers. The columns *month*, *day* are extracted from the column *Date* since they delineate significant information about the season. Similarly, the Latitude, Longitude coordinates are also separately ascribed into different columns. The columns *Sunset_x_hour*, *Sunset_x_min* are derived from the *Sunset_x* column depicting the exact hour and minute at which the sunset take place. The columns *Sunrise_x_hour*, *Sunrise_x_min* are derived from the *Sunrise_x* column depicting the exact hour and minute at which the sunrise take place.

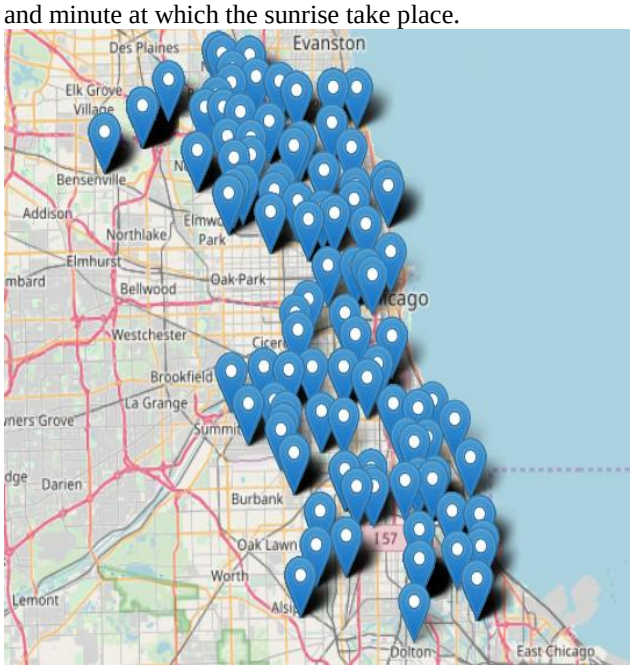

Fig.4 Places in Chicago where the outbreak is seen.

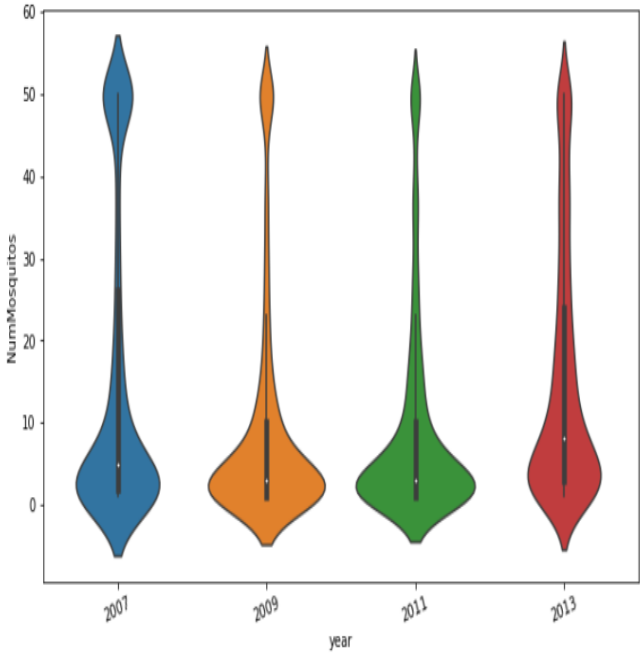

Fig.5 Violin plot representing the number of mosquitoes in the year 2007, 2009, 2011, 2013.

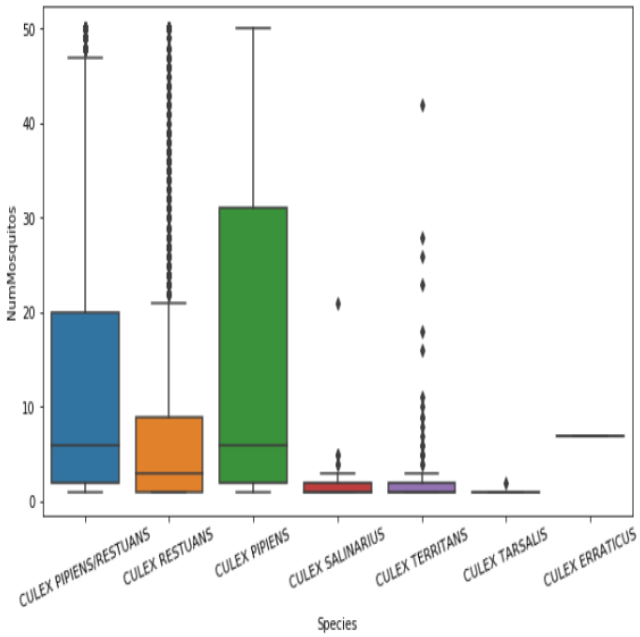

Fig.7 Box plot representing the variation in the mosquitoes count for each species of mosquito.

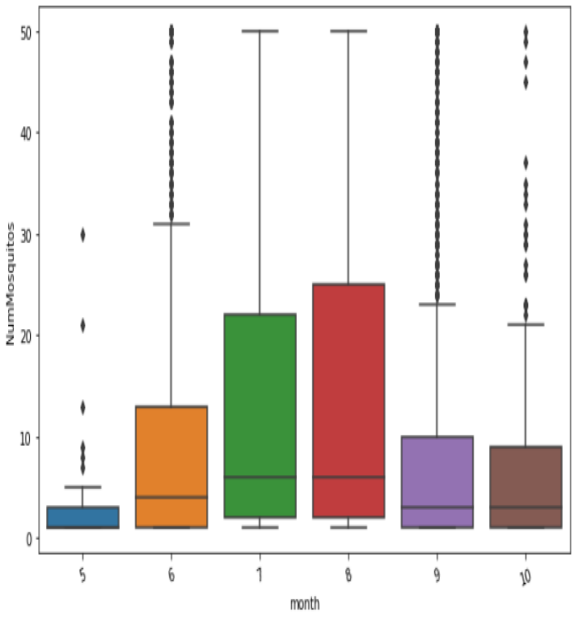

Fig.6 Box plot representing the variation in the mosquitoes count in the months from May to October (Mosquito season).

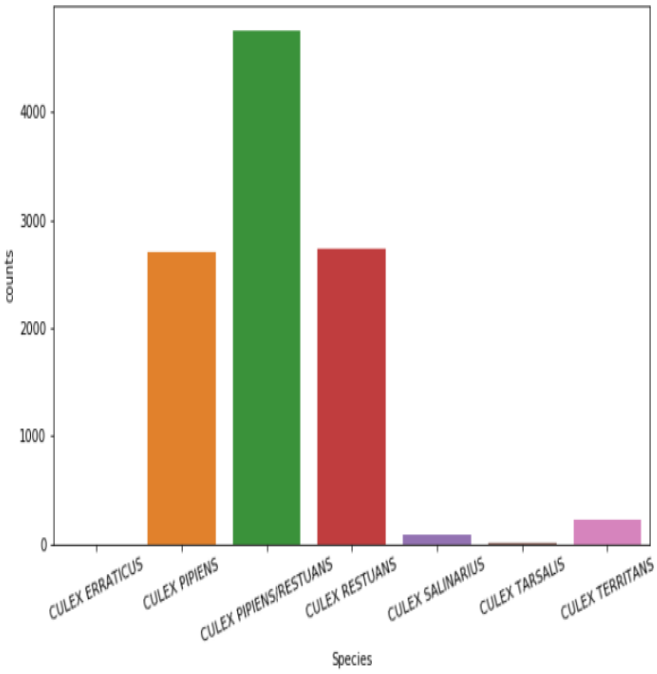

Fig.8 Bar plot representing the mosquitoes count for each species of mosquito.

The outbreak is mainly seen in the northern part of Chicago as shown in Fig.4. The mosquito population is at its peak in the month of August, Number *8* in the X-axis denotes the month August, as shown in Fig.6. The average number of traps found in a locality can go over 20 during its peak. The month of May has very less population of mosquitoes during the season. The population of mosquitoes in each trap collected every time varies over the year as shown in Fig.5. The least number of mosquitoes found is 1 and the maximum number of mosquitoes found is 50. In the year 2007, there were many instances where the trap had more than 40 mosquitoes. However, in the year 2009 and 2011, most of the traps had a minimum population.

The most commonly found mosquito species is *CULEX PIPENS*, which has a higher average trap count in the city as shown in Fig.7. *CULEX TARSALIS* has the least average trap count of all the species of mosquitoes found in the city. *CULEX PIPENS/RESTUANS* has the highest total population in the city with 4752 mosquitoes belonging to the breed as shown in Fig.6. *CULEX ERRATICUS* has the lowest total population count in the city with just 1 mosquito belonging to the breed being found in the city as shown in the Fig.8.

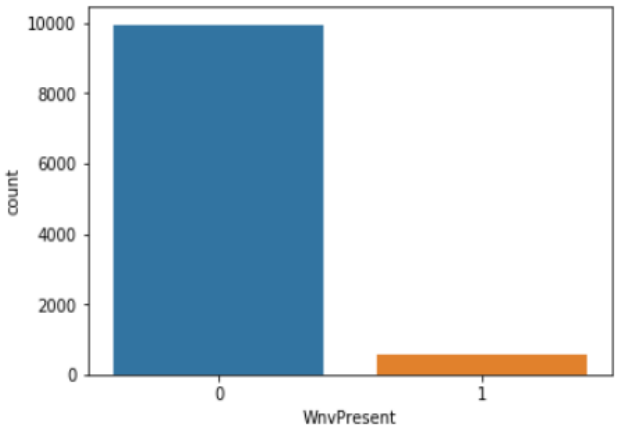

Fig.9 Bar graph indicating the number of instances in the dataset where the virus is present and the number of times they are absent.

The dataset is imbalanced as the number of instances where the Virus is absent dominates the instances where the virus is present as shown in Fig.9. The number of instances in the dataset where the virus is absent is 9955 and the number of instances where the virus is present is 551. The ratio is almost 18:1 (18.06:1 to be exact). Therefore, the dataset is balanced using upsampling where the number of instances of the presence of virus is increased to 9955 and thus making the ratio 1:1. The upsampling is preferred over downsampling because the dataset is in the form of discrete time series and down sampling will remove the important instances from the dataset. The oversampling method used in the present work is Synthetic Minority Over sampling Technique (SMOTE). The training and test data is split in the ratio of 70:30 from the train.csv dataset.

IV. SYNTHETIC MINORITY OVER SAMPLING TECHNIQUE (SMOTE)

SMOTE is an oversampling technique that samples by adding synthetic data over the existing the dataset [5]. Advantages of SMOTE over sampling with replacement technique is that it preserves the existing instances that are vital to the dataset. Synthetic examples are produced by manipulating the feature space without disturbing the data. The oversampling over the minority class is performed by taking each sample belonging to the minority class and building synthetic samples over it such that the synthetic samples stay closer to the line segment joining the minority class' nearest neighbors. The difference between the sample and its associated neighbor is taken to produce the synthetic sample. Then the difference is multiplied with a random value in the range 0 to 1 and is added to the sample vector. Thus, it effectively ensures that the selected random point is between the line segment joining 2 feature vectors [6]. The set of synthetic samples allows the decision regions to be larger and less specific to certain data points.

V. MACHINE LEARNING TECHNIQUES

A. Support Vector Machine (SVM)

SVMs are trained with learning algorithms from optimization theory and uses linear function in a high dimensional feature space [7], [8]. It is in this dimensional space they plot the vectors and finally build the hyperplane which separates opposite classes. SVMs are based on structural risk minimization.

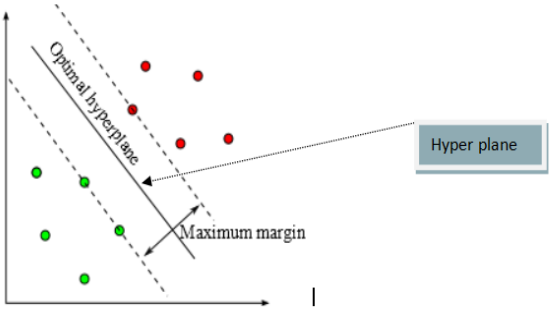

Fig.10 Graphical representation of hyperplane and the support vectors.

The most important aspect of SVM is to find the right hyperplane [9]. Two parallel hyperplanes are constructed on either side of the data separating hyperplane as shown in Fig.10. The hyper plane ensures that the model does not over-fit on a particular sample of data and addition of new data can be easily classified. Parallel hyperplanes are constructed beside the classifying hyperplane such that distance between separating hyperplane and the parallel hyperplanes is maximum. SVMs use kernels, which are mapping functions, to transform an input data from one form to another. The functions are - linear, non-linear, polynomial, sigmoid and Radial Basis Function (RBF). In the present work RBF is used as the SVM kernel.

B. Random Forest

Random forest is a supervised learning algorithm that consists of an ensemble of decision trees to solve classification and regression tasks [10], [11]. It outputs the class which is the mode of the classes (classification) and mean of the predicted values (regression). Random forests are known to outperform shallow neural networks in many cases.

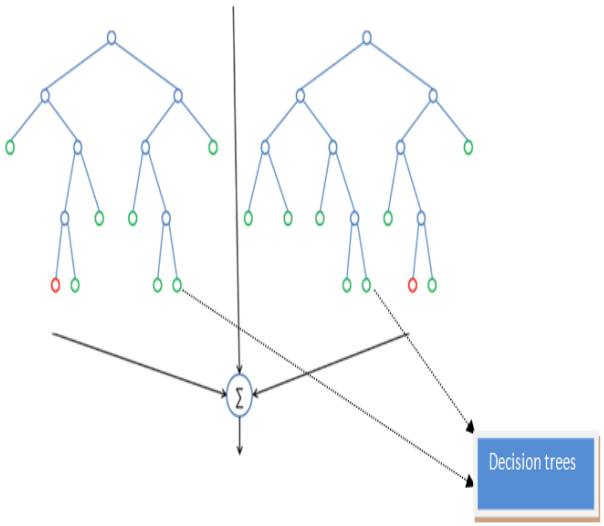

Fig.11 Overview of the two decision trees that constitute a Random forest.

Fig.11 depicts the outputs from two decision trees to produce the hypothesized output.

C. Boosting classifier

Boosting is an ensemble algorithm used to transform an ensemble of weak learners into a set of strong learners [12], [13]. It reduces the variance of the model and ensures that the model does not overfit. Initially, a model is created on the training data and then this model is altered to give a corrected second model. Models are added until the training data is perfectly predicted or if the model limit has been reached. The boosting algorithms used in the present work are – gradient boosting and XGB.

1) Gradient Boosting

Gradient boosting is a machine learning technique which produces a model (ensemble of weak model) and trains those weak models in an sequential manner [14]. The algorithm uses gradients of loss function to sequentially find the weaknesses of the model and the drawbacks associated with it. The loss function is a measure that depicts how efficient are the model's coefficients at fitting the data [15]. An added advantage of the gradient boosting is that it allows the users to control the cost function instead of the loss function as the loss function offers very little control over the model.

2) XGBoost

XGBoost is an implementation of gradient boosting algorithm. The algorithm provides good computational efficiency and is faster when compared to the other implementations of the gradient boosting algorithms [16]. It is similar to gradient boosting technique apart from the fact that it performs an additional custom regularization in the objective function to ensure that training data does not overfit.

VI. RESULTS

Table.1 The models and its F-1 scores are tabulated.

Algorithm	Accuracy (in %)
Gradient Boosting	92.71
XGB	**92.93**
SVM	91.16
Random forest	92.78

All the 4 models are trained and tested with training set and test set, respectively. The results are tabulated as shown in Table.1. Out of the 4 models, XGB performs better. Its F-1 score (92.93) is marginally better than that of Random forest, which gave a score of 92.78. The gradient boosting algorithm gave a score of 92.71 whereas the SVM algorithm gave a score of 91.16.

VII. CONCLUSION

The West Nile River Virus prediction was discussed in this paper and the dataset from the Kaggle repository was used to predict the presence of the Virus. Instead of the traditional method, which involves checking the presence of virus in the traps of mosquitoes belonging to a particular locality at regular intervals, past set of data can be used to predict whether the locality is being affected by the epidemic and whether the area needs to be sprayed with the pesticides. Among the models used, The XGB model gave the best performance with F-1 score of 92.93 closely followed by Random forest at 92.78 and gradient boosting at 92.71. The ensemble classifiers performed better when compared with SVM (91.16). The work can be extended to different cities and countries by collecting the data related to the locality, which also includes the weather related data.

REFERENCES

1. Mitch Campion, Calvin Bina , Martin Pozniak, Todd Hanson, Jefferson Vaughan, Joseph Mehus, Scott Hanson, Laura Cronquist, Michelle Feist, Prakash Ranganathan, Naima Kaabouch and Mark Boetel, 2016, "Predicting West Nile Virus (WNV) occurrences in North Dakota using data mining techniques", 310-317. 10.1109/FTC.2016.7821628.
2. Eliza Little, Scott R. Campbell and Jeffrey Shaman, 2016, "Development and validation of a climate-based ensemble prediction model for West Nile Virus infection rates in Culex mosquitoes, Suffolk County, New York", Parasites & Vectors. 9. 443. 10.1186/s13071-016-1720-1.
3. S. Agarwal, "Data Mining: Data Mining Concepts and Techniques," *2013 International Conference on Machine Intelligence and Research Advancement*, Katra, 2013, pp. 203-207.
4. Goutte, Cyril & Gaussier, Eric, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation", Lecture Notes in Computer Science. 3408. 345-359. 10.1007/978-3-540-31865-1_25, 2012
5. Nitesh Chawla, Kevin Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer, 2002, "SMOTE: Synthetic Minority Over-sampling Technique", J. Artif. Intell. Res. (JAIR), 16. 321-357. 10.1613/jair.953.
6. Alberto Fernández, Salvador García, Francisco Herrera and Nitesh V. Chawla. "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary." *J. Artif. Intell. Res.* 61 (2018): 863-905.
7. M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18-28, July-Aug. 1998.
8. Theodoros Evgeniou and Massimiliano Pontil, 2001, "Support Vector Machines: Theory and Applications" 2049. 249-257. 10.1007/3-540-44673-7_12.
9. Yujun Yang, Jianping Li and Yimei Yang, "The research of the fast SVM classifier method," *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Chengdu, 2015, pp. 121-124.
10. L. Breiman, "Random forests", *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
11. T. K. Ho, "Random decision forests", *Proceedings of the Third InternationalConference on Document Analysis and Recognition*, pp. 278-282, 1995.
12. Bühlmann, Peter, "Bagging, Boosting and Ensemble Methods", Handbook of Computational Statistics. 10.1007/978-3-642-21551-3_33, 2012
13. T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano, "Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data," in *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no. 3, pp. 552-568, May 2011.
14. T. G. Dietterich, G. Hao, A. Ashenfelter, "*Gradient Tree Boosting for Training Conditional Random Fields*".
15. J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *Annals of Statistics*, 2000.
16. Tianqi Chen & Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System". 785-794. 10.1145/2939672.2939785, 2016

AUTHORS PROFILE

Rishickesh Ramesh is working toward B.Tech degree in the Department of Information Technology, SSN College of Engineering. His research interests include machine learning techniques for natural language processing and understanding, data analytics, Internet of Things and deep learning.

A. Shahina is a professor in the department of Information Technology at SSN. She has 14 years of teaching and research experience, with over 5 years of research exclusively in the field of Speech Processing, one of the widely growing and popular research areas. She aims to develop speech based clinical applications, and technologies for viable biometric person authentication systems through sustained research. Her areas of interest include machine learning, deep learning, and speech processing.

Dr. A. Nayeemulla Khan is the Dean Academics and Professor in the School of Computing Science and Engineering at VIT Chennai. He was previously associated with the Airports Authority of India as a Senior Manager ATC and as a Research Scientist at Acusis Software India Pvt. Ltd. His areas of interest include speech and speaker recognition, machine learning, brain computer interface among others.