Avian Influenza Prediction Using Machine Learning Approaches: A Review

Maana Shori, Kriti Saroha

Centre for Development of Advanced Computing, Noida, India

Corresponding author: Maana Shori, Email: maanashori945@gmail.com

The avian influenza virus can be a cause of economic devastation due to its impact on poultry and potentially on the economy, making it the cause for a potential pandemic. By predicting the disease, the consequences can be mitigated, so the work done in this area so far and the gaps in the study are reviewed in this paper. An approach is also proposed in this direction.

Keywords: Machine learning, Avian Influenza, Outbreak, Ensemble Classifiers.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 177–184. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-18

1 Introduction

The Hong Kong SAR event in 1997 validated the pandemic potential of Avian Influenza (H5N1) and brought new insight into how a new pandemic virus could emerge. Previous to 1997, pigs were considered the mixing vessels for virus re-assortment because of the fact that pigs possessed receptors on the cells of their respiratory tract, for each avian and human influenza viruses. However, from the Hong Kong SAR event it could be concluded that since human beings could also be infected with the avian influenza virus, they too could serve as mixing vessels and act as vessels for the virus genes' exchange. Through these findings, H5N1 was then seen to have pandemic potential.

Highly pathogenic avian influenza (HPAI), formerly called fowl plague, is known to cause a broad set of symptoms ranging from being a mild illness to being highly contagious, and once introduced into the domestic poultry population has the potential to spread widely through contaminated insentient objects present in the environment, leading to a wide and swift spread of the disease. Thus domestic poultry is extremely vulnerable to avian influenza where the mortality rates can reach up to 100%. Some of the symptoms in the infected poultry may include nervous system disorders, sneezing, diarrhea, coughing, edema of the head and sudden death.

Researchers have made efforts in the controlling and eradication of the HPAI disease, but the repeated losses to poultry have been a continuous threat to human lives. The different factors that contribute to the spread of the disease can be used to determine the establishment and the impact of the avian influenza virus outbreaks in the future. Research on many environmental conditions such as land cover, agricultural factors, trading activities, poultry population and their farming have found these to be important factors causing the introduction and spread of the disease occurrence [1, 2, 5, 6].

The rest of the paper is arranged in the following sections; **Section 2** presents the literature survey related to the topic and also gives a comparative analysis of the related work, **Section 3** explains the proposed methodology and the steps, which would be taken to implement the research work. **Section 4** gives the conclusion and future work followed by the list of references.

2 Literature Review

The brief introduction along with the background into Avian Influenza disease has been discussed in section 1. Avian Influenza outbreaks can start anytime anywhere due to the migratory birds, so the prediction of this pandemic is very essential. The literature survey presents a discussion on the related work in this area.

Yousefinaghani et al.[3] mapped, executed and verified a decision support framework that predicted and monitored the occurrence of the avian flu events. The authors collected data from Twitter, preprocessed the data and extracted the rules and facts from it to form a knowledge base. Using the knowledge base questions about the degree of risk at various geographical places were responded to. The authors concluded that through their proposed framework prediction of high pathogenic viruses could be done more accurately as compared to low pathogenic viruses.

Painuli et al. [7] used existing COVID-19 data to generate an estimate about the number of positive cases in the future. The authors used several machine learning approaches and in their research discussed the ones that had the best accuracies with the aim to predict the possibility of being infected by the virus and forecast the number of positive cases as well.

Venkatesh et al.[8] discussed the prediction of disease based on some given symptoms and created a system using Machine learning algorithms namely, Decision Tree, Support Vector Machine (SVM),

Random forest, KNN and Naive Bayes classifier. The dataset which is used by the authors included records of several patients who were individually diagnosed with 41 different diseases.

PillaSrinivas et al. [1] applied the Bayesian classifier to predict the severity of patient condition due to Swine Flu, which is a type of Influenza virus. The authors use 14 sample records of different suspected patients and use these sample reports for creating the dataset. The Naive Bayesian Classifier was used, which does not need large quantities of data to build the training set, and so was able to produce a fast outcome leading to rapid identification of the disease, which enables early treatment.

Chauhan et al.[9] analyzed the performance of various classification algorithms namely, XGBoost, Logistic Regression, Decision Tree, KNN, Random Forest, Naive Bayes and SVM. The algorithms were applied to a dataset obtained from UCI to find the most accurate algorithm that could predict a patient's chances of developing heart disease.

Taj et al. [14] discussed the most widely used machine learning (ML) and deep learning (DL) models for understanding COVID-19 behavior by investigating time series data. The authors proposed a model for examining and forecasting COVID-19 by regional distribution.

Tapaket al. [10] discussed the prediction of influenza-like illnesses and compared the accuracies obtained by applying different machine learning approaches namely, Random-forest, Artificial Neural-Network (ANN), and SVM. The study was carried out using a dataset that included weekly influenza cases from Iran.

Naiyar et al. [11] forecasted the number of negative cases or the positive cases of dengue outbreak based on seven attributes and seven machine learning algorithms. It was concluded that LogitBoost ensemble model was the topmost performance classifier technique that had reached a classification accuracy of 92%.

Kalipe et al.[12] used meteorological data of malarial cases, in contemplation of examining and forecasting the occurrence of malaria. The authors used various classifiers namely, Extreme Gradient Boost (XGBoost), Random Forest, SVM, Logistic Regression, Naive Bayes, KNN and ANN. The authors concluded that meteorological data could be used to predict malarial outbreaks which can result in the saving of lives lost due to the disease. XGBoost algorithm was found to be particularly efficient for their study.

Agrawal et al. [13] used ensemble and simple classifiers to predict patients' chances of liver, heart and diabetes diseases. The dataset for the study was taken from the University of California, Irvine's website. The authors obtained accuracies for all three datasets by applying the machine learning model. The best accuracy was obtained for the liver dataset. Also, by reducing the number of features, the authors got somewhat reduced accuracies but were all well within the acceptable ranges.

Singh et al.[2] used different machine learning models namely, Random Forest, Gradient Boosting Machine and Support Vector Regression to predict influenza. The authors improved their forecast by including meteorological factors like precipitation, temperature and humidity which resulted in improved accuracy for forecasting influenza.

Biswas et al.[6] studied the contribution of climatic factors in Bangladesh, concerning the incidence of Highly Pathogenic Avian Influenza outbreaks. The authors, in their study, have used the ARIMA and SARIMA models for obtaining the relation of outbreak occurrences due to particular meteorological factors like cloud cover and average rainfall.

Kane et al.[15] analyzed the time-series structure of outbreak intensity for highly pathogenic avian influenza (H_5N_1) for the country of Egypt using ARIMA and Random Forest time series models.

Herrick et al.[4] used the Random Forests classifier to develop a predictive plot for depicting Avian

Influenza at a global scale. The authors described predictors and environmental factors that may be a cause for the spread and infection of highly pathogenic avian influenza in wild birds. The authors have identified the highest risk of the outbreak to be the northern regions in their study.

Table 1 gives a comparative analysis of the works discussed so far. The 12 papers discussed above, used different machine learning classifiers for the identification, evaluation and prediction of diseases. Most papers applied Random Forest, Naive Bayes, and Support Vector Machine, for forecasting different diseases. ARIMA and LSTM were also employed that included cases that were observed over some time, for making predictions for the number of cases in the future. It is observed that the comparison of the simple and ensemble classifiers for the prediction of Avian Influenza disease, in particular, has not been done before, to the best of my knowledge. There has been no study that compares how significant the difference in accuracy is between ensemble and individual classifiers. Also, for several parts of the Asian region, the study for the identification and evaluation of H5N1 has not been carried out using data that includes data on common backyard poultry like ducks, geese, chickens, etc. Since ducks, geese, chickens come in direct contact with humans, it is important to include their population for the prediction. Comparison of accuracies obtained before and after preprocessing of data rendered different results in [10]. Also, a change in values of accuracies with the number of features being considered was carried out in [11]. Thus, the work may be extended and data in future work may also include attributes/ features on common backyard poultry like ducks, geese, chickens, etc.

Table 1. A comparative study of related works on disease prediction using ML models

Paper Name	**Algorithms Used**	**Accuracy obtained/ Result**
Forecast and prediction of COVID-using machine learning [7]	Random Forest Extra Tree Classifier	91.63% 93.62%
A framework for the risk prediction avian influenza occurrence: An Indonesian case study [3]	Rule discovery algorithms	The effect of each rule and its risk were computed
An Artificial Intelligent based System for Efficient Swine Flu Prediction using Naive Bayesian Classifier [1]	Naive Bayes	For new set of symptoms the patient's diagnosis was done, and the prediction of whether the patient has the disease or not was made
Comparative evaluation of time models for predicting influenza outbreaks: application of influenza-like illness data from sentinel sites of healthcare centers in Iran [10]	SVM ANN Random Forest	89.2% 88.9% 86.2%
Machine Learning for Dengue Outbreak Prediction: A Performance Evaluation of Different Prominent Classifiers [11]	Naive Bayes Decision Tree ANN SMO Logistic Regression kNN LogitBoost	81% 84% 81% 80% 85% 75% 92%
Predicting Malarial Outbreak using Machine Learning and Deep	kNN SVM	86.21% 92.69%

Approach: A Review and Analysis [12]	ANN Naive Bayes Random Forest XGboost Logistic Regression	25.83% 91.69% 93.94% 96.26% 92.44%
Disease Prediction Using Machine Learning [13]	Support Vector Machine	Heart Dataset- 75.4%, DiabetesDataset- 78.6% Liver Dataset- 75.9%
Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks [15]	Random Forest, ARIMA	Random Forest model's Mean Squared Error (MSE), was less than both the ARIMA's simulated retrospective MSE and the prospective ARIMA's MSE
Modeling and Roles of Factors in Outbreaks of Highly Pathogenic Avian Influenza H5N1 [6]	ARIMA, SARIMA	ARIMA (1,0,1): AIC=159.16 Error%= 29.7%
Towards Using Recurrent Neural Networks for Predicting Influenza-like Illness: Case Study of Covid-19 in Morocco [14]	Long Short Term Memory	Epidemic prediction and study to make predictions on Covid-19 pandemic growth in Morocco, using LSTM model was done
Identification of Disease Prediction Based on Symptoms Using Machine Learning [8]	Random Forest SVM Decision Tree KNN Naive Bayes	94.12% 95.01% 95.13% 92.68% 94.25%
Disease Prediction using Machine Learning [9]	Naive Bayes Decision Tree Random Forest	92.90% 93.85% 97.64%
Ensemble Approach for Zoonotic Disease Prediction Using Machine Learning Techniques [2]	Gradient Boosting Machine + Random Forest Random Forest + Support Vector	44.6% 67.2%
A global model of avian influenza prediction in wild birds: the importance of northern regions [4]	Random Forest	79%

3 Proposed Approach

The prediction of a future outbreak of HPAI would be made. Fig. 1 shows the steps that would be taken to achieve the objective. Firstly, the dataset would be collected and the initial pre-processing would be done to ensure that no features were overlooked. The step of feature extraction would involve the scaling and translation of the data if required. The sample rows that contain missing values may be dropped.

The data would then be divided into training and testing sets, randomly. The training data would be used to train the model while the test dataset would be used in the end to assess the performance of the final model. The test data would be used only once to avoid over fitting. To enable the reuse of the test dataset for the model evaluation, a validation dataset may also be created. The main purpose of the two approaches is to increase computational efficiency, avoid over fitting and remove noise.

Also, feature selection and dimensionality reduction will be undertaken to drop those attributes from the dataset that are not so important and study the effect on the performance of the classifiers. For analyzing the features, a heat map would be used. According to the heat map, the most important features in predicting future outbreaks will be retained. After the dataset is cleaned and analyzed, various machine learning models, both simple and ensemble, would be applied. After that, the performance validation for both types of classifiers would be done. Finally, out of these, the best classifier would be determined based on the comparison done between the performance validations obtained for both types of classifiers.

There will be now 2 datasets, one containing all the attributes and another with reduced number of attributes. The dataset with reduced number of attributes would also be divided into training and testing sets, randomly. The datasets obtained thereafter would be applied to different classifiers, in this case separately to simple and then ensemble classifiers. Performance evaluation would be done to evaluate the models' performances on complete set of attributes and reduced attributes. A confusion matrix would be constructed to obtain the accuracy, sensitivity, specificity, recall and precision of each model. The model with the best accuracy would be considered for the purpose of prediction of the avian influenza disease.

4 Conclusion

Mutation events and genetic re-assortment may be increased due to the rapid spread of the disease. This may make the spread uncontrollable which is what makes the spread of the avian influenza a threat that could be realized as a pandemic in the near future. For the prevention of the devastation caused by the disease, forecasting the disease could be a step in the right direction.

For proper disease management and control, timely predicting the potential disease outbreak plays a key role so as to reduce the devastating effects caused by it both in terms of loss oflives and economy. Avian Flu may present a potential challenge by hampering the development of a particular region and possibly the world, much similar to the current COVID-19 pandemic.

In the proposed approach, various classifiers would be applied and the accuracies obtained would be compared to determine the one that is most efficient at predicting the Avian Influenza in Asian Countries (majorly India). The dataset used would include maximum features that impact the occurrence of Avian Influenza, such as meteorological factors and duck and chicken populations. In previously done research, chickens and pigs, as important transmitter of avian influenza, have been studied widely. So, for this research, duck and geese populations will also be considered for the prediction of avian influenza and its spread in the Asian region. Results using classifiers other than the ones used in earlier studies will also be explored to predict avian influenza disease. Several different climatic variables will also be considered as attributes to study the refinement of the predictability of the proposed model and also to verify if the impact of the climatic variables is in accordance or in contradiction to prior research.

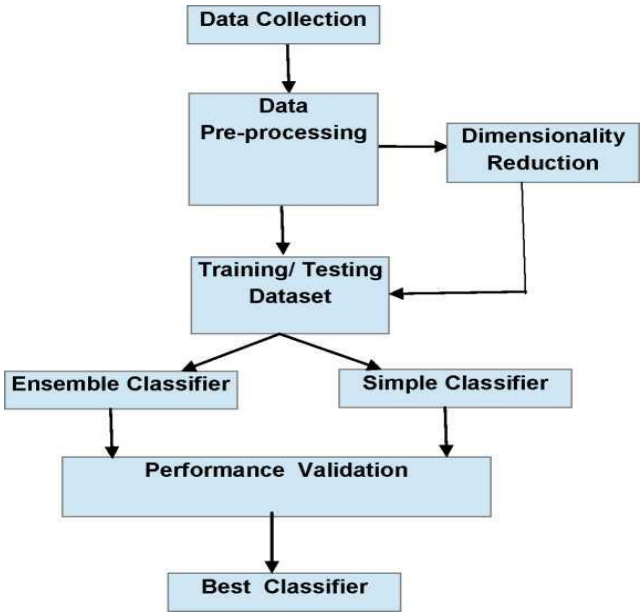

Fig. 1. Block Diagram of the proposed methodology

References

[1] Pilla, S., Bhattacharyya, D. and Chakkaravarthy, D. M. (2020). An Artificial Intelligent based System for Efficient Swine Flu Prediction using Naive Bayesian Classifier. International Journal of Current Research and Review, 12:134-139.

[2] Singh, R. K. and Sharma, V. (2015). Ensemble Approach for Zoonotic Disease Prediction Using Machine Learning Techniques. International journal of business, 3.

[3] Yousefinaghani, S., Dara, R., Pojk, Z., Song, F. and Sharif, S. (2021). A framework for the risk prediction of avian influenza occurrence: An Indonesian case study, Public Library of Science One, 16(1).

[4] Herrick, K. A., Huettmann, F. and Lindgren, M. A. (2013). A global model of avian influenza prediction in wild birds: the importance of northern regions, Veterinary Research, Springer, 44:42.

[5] Pandita, A., Yadav, S., Vashisht, S. and Tyagi, A. (2021). Review Paper on Prediction of Heart Disease using Machine Learning Algorithms. International Journal for Research in Applied Science and Engineering Technology, 9:2937-2940.

[6] Biswas, P. K., Islam, M. Z., Debnath, N. C. and Yamage, M. (2014). Modeling and roles of meteorological factors in outbreaks of highly pathogenic avian influenza H5N1. Public Library of Science One, 9(6).

[7] Painuli, D., Mishra, D., Bhardwaj, S. and Aggarwal, M. (2021). Forecast and prediction of COVID-19 using machine learning. Data Science for COVID-19, 381-397.

[8] Venkatesh K., Dhyanesh K., Prathyusha M., Teja C.H.N (2021). Identification of Disease Prediction Based on Symptoms Using Machine Learning. JAC : A Journal Of Composition Theory, 14(6).

[9] Chauhan, R. H., Naik, D. N., Halpati, R. A., Patel S. J. and Prajapati, A. D. (2020). Disease Prediction using Machine Learning. International Research Journal of Engineering and Technology, 7(5).

[10] Tapak, L., Hamidi, O., Fathian, M. and Karami, M. (2019). Comparative evaluation of time series models for predicting influenza outbreaks: Application of influenza-like illness data from sentinel sites of healthcare

centers in Iran. BMC Research Notes, 12.

[11] Iqbal, N. and Islam, M. (2019). Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers. Informatica, 43.

[12] Kalipe, G. and Gautham, V. and Behera, R. (2018). Predicting Malarial Outbreak using Machine Learning and Deep Learning Approach: A Review and Analysis. International Conference on Information Technology, 33-38.

[13] Agrawal, A., Agrawal, H., Mittal, S. and Sharma, M. (2018). Disease Prediction Using Machine Learning. 3rd International Conference on Internet of Things and Connected Technologies.

[14] Moulay, T., Rachida, Ait, E. M., Zakariyaa, Jakimi, A. and Hajar, M. (2020). Towards Using Recurrent Neural Networks for Predicting Influenza-like Illness: Case Study of Covid-19 in Morocco. International Journal of Advanced Trends in Computer Science and Engineering, 9: 7945-7950.

[15] Kane, M., Price, N., Scotch, M. and Rabinowitz, P. (2014). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC bioinformatics, 15:276.