Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 11s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Effective identification of Swine flu-H1N1 virus using machine learning algorithms

Nashreen Begum J[1], Dr. Sandeep Chahal[2]

[1]Research Scholar, NIILM University, Kaithal, Haryana
Email ID: jnashreen1983@gmail.com
[2]Associate Professor, NIILM University, Kaithal, Haryana
Email ID: sandeepchahal.5@gmail.com

Cite this paper as: Nashreen Begum J, Dr. Sandeep Chahal, (2025) Effective identification of Swine flu-H1N1 virus using machine learning algorithms. *Journal of Neonatal Surgery*, 14 (11s), 426-433.

ABSTRACT

The swine flu is a highly contagious virus and poses a significant threat to human health. The virus affects various physiological systems and presents severe symptoms that harm the affected and the community. Due to its high transmission rate and adverse health impacts, it is imperative that the authorities find ways to detect and predict its outbreak early. Therefore, machine learning presents an avenue to help identify the illness before it gets out of hand. Review of the use and ways through which machine learning can be used will lead to the ability of predicting the infection with precision and saving thousands of lives. Current research work proposed the type of swine flu virus infecting a patient, using Machine Learning techniques. Developed in the prototype application are different ML techniques like Support Vector Machine, logistic regression, Decision Tree and Naive Bayes. The process begins by pre-processing the data to clean the input from possibly noisy values, missing data, scale, and data reduction so that the input data will be prepared and ready for analysis. Subsequently, the pre-processed data is fed into the ML algorithms in order to elicit predictions regarding the type of swine flu virus from the patient's records. Other measures adopted in this research, for example, accuracy, precision, recall, F1-score, were also used in the performance evaluation of the ML models. Such measures will be essential in reflecting the effectiveness of each single model in the right classification of the swine flu virus type. The results show that the Support Vector Machine have outperformed the other ML models in all evaluation metrics, scoring an accuracy rate of 96.32%. This implies that the ensemble techniques are very effective in identifying the type of swine flu virus present in the patient to take more informed medical decisions and treatment strategies.

Keywords: *Machine Learning, Support Vector Machine, logistic regression, Decision Tree and Naive Bayes.*

1. INTRODUCTION

The swine flu virus is extremely communicable along with has the ability to quickly propagate through a variety of vectors, including the air and water. A kind of the virus that causes influenza that represents a severe hazard to public health all over the world is the particular strain in question. The seriousness of the illness is highlighted by the fact that viruses that cause outbreaks, whether they are local or global, frequently result in mortality [1]. It has been reported that the H1N1 virus has caused human infections in a number of different countries, which highlights the worldwide effects of the virus. Symptoms of the swine influenza in humans can range from moderate to severe respiratory diseases and other consequences. They are especially susceptible to the respiratory tract, which is particularly susceptible to the disease. According to estimates provided by the World Health Organisation (WHO), the influenza virus is responsible for the deaths of between 300,000 and 600,000 people each year, affecting anywhere from 10 to 25% of the world's inhabitants [2]. The economic repercussions of these infections are large, in addition to the fact that they cause a significant amount of morbidity.

The financial impact of influenza-related diseases is tremendous, estimated to be between $75 billion to $190 billion yearly in the United States alone. This is only the United States. Not only does this financial burden have an impact on the medical sector, but it also contributes to broader consequences for the economy of the entire world. Swine flu outbreaks have the potential to put a strain on healthcare resources, cause disruptions in production, and result in enormous losses for organisations and governments all over the world. In order to lessen the impact that the swine flu has on both general health and economic activity, it is necessary to develop and implement efficient measures regarding the mitigation, early identification, and treatment of the disease [3]. The purpose of this research is to suggest the utilisation of a method that is based on machine learning for the classification and forecasting of epidemics of swine flu. The algorithm that is being used

is designed to make predictions about the potential spread of the sickness, and it also makes adjustments to those forecasts on a continuous basis as new instances become known. The capability of the technique to construct exclusion zones, which are determined by a set of limits that are produced by computational processes, is one of the most important features of the technique [4].

The study and investigation of trends within information sets, especially those connected to outbreaks of swine flu, can largely benefit from the application of machine learning, which offers substantial advantages. The algorithm is able to successfully interpret information gathered from datasets used as training and provide recommendations that utilise the patterns that it has learned by utilising classifiers that depend on machine learning techniques [5]. The approach entails training the classifiers using labelled datasets, in which every point of data has a connection with a known result (for example, whether the individual is diseased or not infected). The information obtained from the training data is subsequently utilised by these neural networks in order to derive predictions from data that has not yet been seen. Through the utilisation of this predictive skill, the algorithm is able to foresee epidemics of swine flu and classify individuals according to the likelihood that they would become infected. Overall, the incorporation of machine learning approaches into the classification and prediction of swine flu shows promise for enhancing our comprehension of the dynamics of the disease and allowing the development of more efficient response tactics [6]. Healthcare workers are able to better forecast and control epidemics of swine influenza by using the potential of machine learning, which ultimately leads to superior outcomes for the general population.

A new method that makes use of machine learning techniques for the purpose of providing real-time forecasts of the occurrence of swine flu is presented in this line of research. Symptoms of swine flu, which is referred to as H1N1 influenza, include throat pain, chills, weakness, vomiting, runny nose, body pains, coughing, and fever. Swine flu is frequently referred to as H1N1 influenza. It is possible that some cases will exhibit multiple instances of the aforementioned signs, with breathing being the most prevalent of the illnesses [7]. On the basis of the presentation of symptoms, machine learning classifiers are an essential component in the process of forecasting the state of swine flu infection. The Support Vector Machine, logistic regression, random forest, Decision Tree, Naive Bayes, and Ensemble methods are all examples of a classifier that falls under this category. Considering the inherent difficulty of the task, these classification algorithms are capable of delivering binary outcomes ("Yes" or "No") concerning the existence of swine flu in people. During the procedure, the classifiers are trained with the help of labelled datasets that contain details regarding symptoms and the accompanying diagnosis for swine flu. The ability to recognise patterns that are suggestive of swine flu infection is developed by the classifiers through the process of learning from these database sets [8]. With this information, they are able to properly determine whether or not a person is likely to have been contaminated with a viral infection according to the symptoms that they are experiencing. Healthcare workers are able to immediately estimate the possibility of swine flu contamination among patients by utilising the predictive capacity of machine learning algorithms. This allows for timely treatments and prevention measures to be taken. This proactive strategy to managing swine flu has the potential to help reduce the propagation of the illness while enhancing the results for patients.

2. LITERATURE SURVEY

The utilisation of data from social media platforms is a promising route for effectively identifying outbreaks of epidemics and sending immediate notifications to the general population. The purpose of this research is to present a comprehensive model that is intended to identify outbreaks of influenza-like illness (ILI) by utilising three essential modules: categorization, visualisation, and prediction through the use of linear regression. In a manner that is analogous, another research endeavour suggested a model for the identification of influenza based on the analysis of tweets [9]. This model utilised Support Vector Machine (SVM) as a method of machine learning for categorization. The study found that the model was effective in reliably recognising tweets connected to flu epidemics, as indicated by the consistent correlation value of 0.89 that was recorded with the highest level of accuracy after the study was conducted. In addition to this, the research included the examination of health-related data on Twitter in order to keep track of health issues in real time [10]. The primary objective of the investigation was to identify tweets that contained references to health organisations. Illness-related tweets were detected and categorised by the utilisation of a domain-tagger named entity technique. However, the study did not take into account tweets from persons who had been diagnosed with the disease.

In addition, the research explored the public's perception of the swine flu and H1N1 in Japan. This was accomplished by utilising regression and statistical methods to examine tweets from a number of places in Japan [11]. The purpose of the study was to identify incidences of influenza-like illness (ILI) in various regions of Japan by systematically identifying tweets that contained the hashtag #flu. In the study, Natural Language Processing (NLP)-based algorithmic methods for classification were utilised in order to categorise tweet messages that were associated with influenza or influenza-like illness. The use of this method allowed for the precise classification of tweets, which resulted in the acquisition of significant knowledge regarding the frequency of conversations on social networking platforms that were relevant to the flu [12].

For the purpose of recognising illness tasks in tweets, a machine learning approach is utilised. Through the use of the social networking service Twitter, a frequency-based analysis was performed on two diseases, including cancer and influenza. The

purpose of their geographical study in every state of the US was for tracking the spread of epidemics in the various regions of the United States by counting the number of cases of both of these illnesses that were produced in the states itself [13]. For the purpose of conducting an analysis of the location facts, the specifics of the location were derived from the timelines of the users. All of the users who cited the information regarding cancer or the flu were included in the dataset, which also included correct details on the areas of the United States. The technique was unusual in that it provided real-time surveillance and was able to detect illness outbreaks in the states of the United States, such as those involving cancer or influenza. Given that this study did not take into account significant aspects such as affected individuals, feelings regarding the disease, and worrisome circumstances, it appears that the most significant shortcomings of this research are the aforementioned. Similarly, the machine learning-based model was utilised by Support Vector Regression (SVR) in order to estimate the regional-based viral occurrence in the United States. They optimised the SVR model by configuring two datasets: one that utilised supported data from the Centres for Disease Control and Prevention (CDC), and the other that had been constructed on tweets from the United States.

3. METHODOLOGY

3.1 Data Preprocessing

After the raw, unorganised facts have been retrieved, it is a difficult process to retrieve important information from a text corpus that is not organised with any particular structure. In order to normalise the data that is going to be deployed as a feature, the data is transformed by employing the natural language processing preprocessing procedures. This includes the removal of stop words, unique characters, stemming, and tokenization, as well as hyperlinks, @mentions, retweets, punctuation, and URLs. The purpose of this procedure is to deliver the corpus in a format that allows it to be processed effectively in order to improve their output. This is accomplished by removing aspects of the information that are not associated with the corpus. Through the process of tokenization, the text of the tweet is broken up into individual word tokens [14]. On the other hand, the stemming methodology involves reducing the term to its root or base word by utilising the NLTK porter stemmed techniques in Python.2.
The biases in labelling were eliminated by having three different people label a specific group of 6000 tweets after the preliminary processing phase was completed. For a tweet that indicates that a person is infected with dengue or influenza, a label of 1 is assigned, and for all of the other tweets that provide certain details about those particular diseases, a label of 0 is assigned [15]. Following that, the labelling is accepted by means of the amount of agreement between the annotation evaluators.

3.2 Classification Algorithms

3.2.1 Logistic Regression

Logistic regression is a statistical technique that models the logistic curve by employing probability functions and operating on the basis of probability values. The classification technique in question is supervised, which means that it necessitates class labels to be present in the data set. Logistic regression is a technique that is typically utilised for binary categorization tasks. It employs training on labelled data to make predictions regarding whether a consequence is positive or negative, which are represented by the numbers 0 and 1, respectively.

For the construction of the logistic regression (LR) model, the subsequent function is typically utilised:

$$f(a_i) = a_i \cdot w + b \qquad \{1\}$$

In formula (1), the symbol w stands for the weights, which are represented by the formula: $w = (w_1, w_2, w_3, \dots, w_k)^T$. The bias term is denoted by the symbol b, and the equation $w = (w_1, w_2, w_3, \dots, w_k)^T$ is used to calculate the bias term. Logistic regression is a statistical technique that aims to minimise the loss operation, which has been defined as the total number of squared variances between the events that were predicted and those that actually occurred:

$$L = \sum_{i=1}^{n} (f(a_i) - b_i)^2 \qquad \{2\}$$

3.2.2 Decision tree

An example of a framework for machine learning that is both widely used and relatively straightforward is the Decision Tree. Regression as well as classification analysis of data sets are both uses that can be accomplished with its assistance. This model divides the data into various categories by using an assortment of queries as the basis for the partitioning. A decision tree is a type of decision tree that uses a categorization procedure that is similar to it. When the tree is first established, all of the samples are stored at the root. In the course of its operation, the model divides the samples into a number of different groups according to the queries, thereby identifying the characteristic to which the sample i belongs. A recursive process is utilised in order to generate the tree to its final form.

The process of selecting features involves analysing measures that include the Gini index, information entropy, or gain ratio. These metrics measure the impurity or unpredictability of the dataset both before and after the partitioning process. An ideal

partition of the feature set is achieved as the outcome of the decision tree technique's iterative selection of the characteristic and threshold that minimises the impurity or maximises information gain.

3.2.3 Support Vector Machine

A Support Vector Machine, often known as an SVM, is a well-known machine learning model that is utilised for classification and regression tasks. It is well-known for its accuracy in managing several sorts of data, including text and images. Due to the fact that it is both accurate and easy to process, support vector machines (SVM) have found widespread use in the field of medicine, particularly in areas such as a diagnosis, forecasting, and categorization. The support vector machine (SVM) is able to successfully segregate different categories of targets by generating the hyperplane in a space that is multivariate. The most important goal of support vector machines (SVM) is to generate an ideal hyperplane in an iterative manner in order to reduce the number of errors. This is accomplished by establishing a marginal hyperplane of the greatest possible length in order to accurately categorise the information being analysed into two categories.

$$w^T x_i + b = 0 \qquad \{3\}$$

The weight of the vector format which is what defines the orientation of the hyperplane, is denoted by the symbol w in this circumstance. It comprises of weights $w_1, w_2, \dots, w_k$ according to the features. b is the parameter that is used for determining the distance between each hyper plane and the vector of features that was first employed.

$$max \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j b_i b_j a_i^T a_j$$

subject to: $\{4\}$

$$\sum_{i=1}^{n} \alpha_i b_i = 0, \alpha_i > 0, i = 1,2, \dots, n$$

3.2.4 Naïve bayes

When it comes to classification problems, Naïve Bayes is a significant classifier that is commonly utilised in numerous applications. Being straightforward in nature makes it an ideal choice for a wide range of classification problems. Bayes' theorem, in conjunction with the likelihood value, is the foundation upon which it operates. Assigning a label to each and every occurrence of an issue that is represented as an independent characteristic is what the Naïve Bayes technique makes use of. That is to say, it is a system of categorization that operates on the foundation of Baye's theorem and makes use of the premise that the variable in consideration is independent.

Generally, model belongs to bayes theorem are using the posterior probability

P(A|C) from P(A). P(C), P(C|A).

$$P(A|C) = \frac{P(C|A)P(A)}{P(c)} \qquad \{5\}$$

From equation 8,

$P(A|C)$ – $Probability\ of\ posterior\ class\ which\ represent\ (A, target)$

$P(A)$ – $Probability\ of\ prior\ class$

$P(C|A)$ – $likelihood\ means\ the\ probability\ of\ given\ predictor\ class$

$P(C)$ – $probability\ of\ prior\ predictor.$

4. RESULT AND DISCUSSION

4.1 Dataset description

The dataset provided for the Kaggle competition "Prediction of H1N1 Vaccination" aimed to forecast whether people acquired an H1N1 influenza vaccination based on demographic, health-related, and immunization-related factors. It incorporated three key files: the coaching unit functions, that contained anonymized records about participants, like age, schooling, and household measurement; the objective labels document, indicating whether every particular person gained an H1N1 flu pictures; and also, the check unit functions, which mirrored the framework of the coaching unit however lacked the objective variable. Additionally, there has been a submission format document that specified the format for submitting predictions. Contributors had been tasked with the use of the coaching records to increase predictive fashions after which making predictions for the check data. Participants strived to develop models using the training data that could accurately predict if an individual received the H1N1 flu vaccine based on the provided variables, in order to submit their predictions for evaluation.

Through the utilisation of Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) curves, the results have been visualised. These curves offer insights into the performance of various models. Each of these curves, which can be found in Figure 2, Figure 3, Figure 4, and Figure 5, illustrates how well each model performs on the dataset. Based on these

numbers, it is clear that the Support Vector Machine (SVM) model, more precisely the one that is marked as 3.2.3, has proven the best level of accuracy. In terms of forecasting H1N1 flu vaccine, this SVM model achieved an accuracy of over 95.36%, while in terms of predicting seasonal flu vaccination, it reached an accuracy of 96.32%. According to these findings, the Support Vector Machine (SVM) model is superior to other models in terms of its ability to reliably forecast the effects of vaccinations against H1N1 and seasonal flu. As a result, this model offers an intriguing strategy for the task of vaccination prediction.

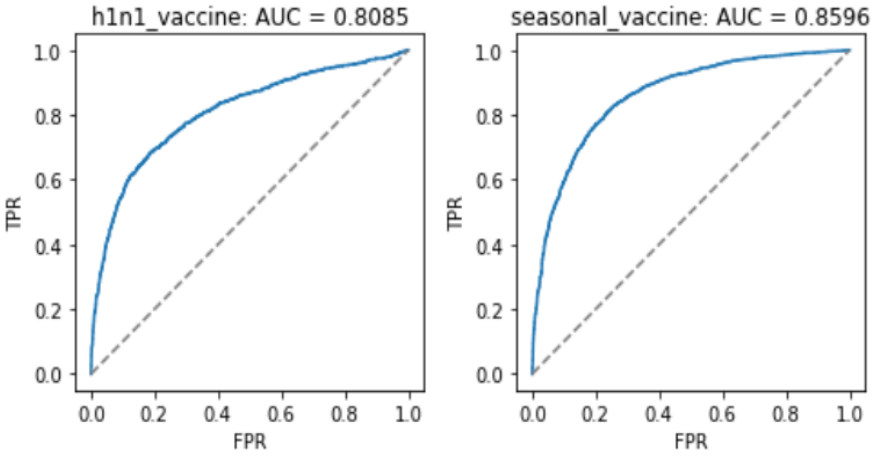

Figure 2. ROC AUC curve for Logistic Regression

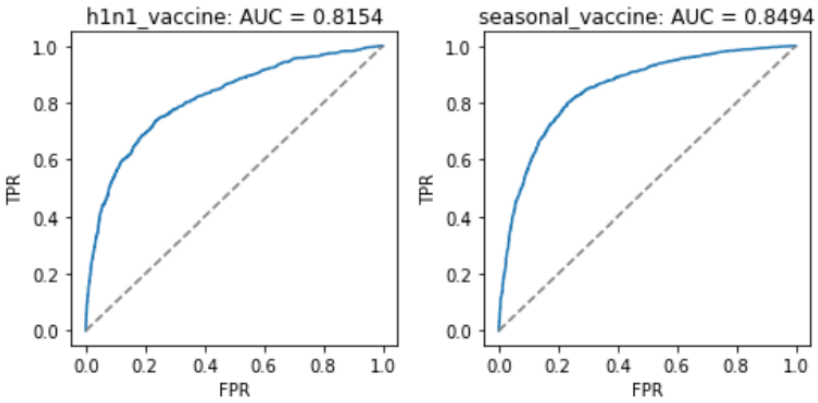

Figure 3. ROC AUC curve for Decision tree

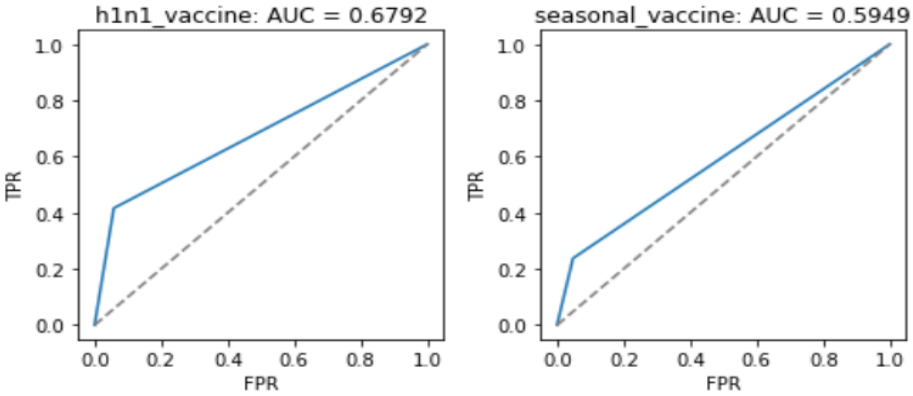

Figure 4. ROC AUC curve for Naïve bayes

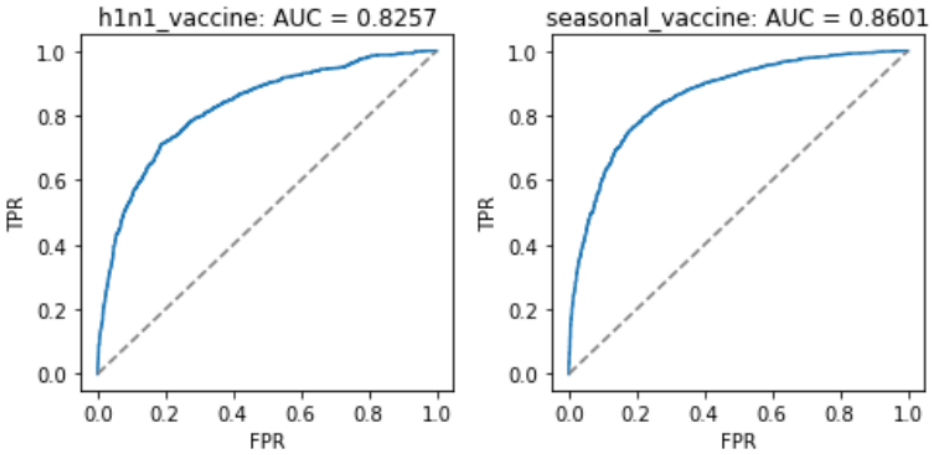

Figure 5. ROC AUC curve for Support Vector Machine

Following **Table 1**, shows the result comparison of machine learning algorithm such as Support Vector Machine, logistic regression, Decision Tree and Naive Bayes by using various evaluation metrices such as accuracy, precision, recall, F1-score, sensitivity, and specificity. In the comparison it is clearly visible that the performance of Support Vector Machine outperforms better than other machine learning approaches used. Also, the comparison figure 6, clearly depicts the performance of all the ML algorithms schematically.

Table 1. Performance comparison of ML algorithms

Classification algorithms	Accuracy	Precision	Recall	F1-score
Logistic Regression	88.38	89.36	87.36	86.26
Decision Tree	91.32	90.36	92.31	91.03
Naïve bayes	93.12	92.31	93.64	90.56
Support Vector Machine	96.32	95.03	94.36	95.06

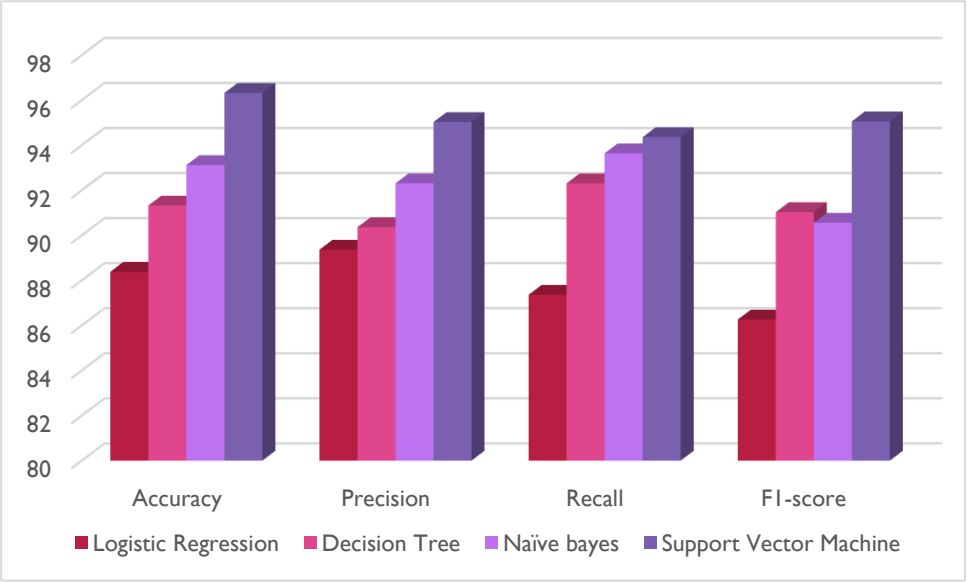

Figure 6. Performance comparison of various ML models

5. CONCLUSION

Because of the highly transmissible nature of the swine flu and the serious health repercussions it can cause, it continues to be an urgent topic of concern. There are promising methods available for early identification and forecasting of outbreaks that can be provided by machine learning, which has the potential to save numerous lives. This research has proved the effectiveness of several machine learning approaches, including Support Vector Machine, logistic regression, Decision Tree, and Naïve Bayes, in determining the type of swine flu virus that is infecting individuals. These approaches were explored through the exploration of various machine learning techniques. A comprehensive evaluation of the effectiveness of all the models was carried out by preprocessing the data and making use of evaluation metrics such as precision, recall, precision, and F1-score. According to the findings, the effectiveness of the Support Vector Machine model was superior to that of other models, as it achieved an excellent accuracy rate of 96.32%. These findings highlight the significance of utilising machine learning in the healthcare industry to improve disease diagnosis and to provide information that can be used to inform medical choice-making, which will eventually lead to healthier outcomes for patients and for the general population.

REFERENCES

[1] Nagaraj, P., AV Krishna Prasad, V. B. Narsimha, and B. Sujatha. "Swine flu Detection and Location using Machine Learning Techniques and GIS." *International Journal of Advanced Computer Science and Applications* 13, no. 9 (2022).

[2] Srinivas, Pilla. "An Evaluation of swine flu (Influenza A H3N2v) virus prediction using data mining and Conventional neural network techniques." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12, no. 4 (2021): 1377-1386.

[3] Srinivas, Pilla, Debnath Bhattacharyya, and Divya MidhunChakkaravarthy. "Prediction of Swine Flu (H1N1) Virus Using Data Mining and Convolutional Neural Network Techniques." In *Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 2*, pp. 557-573. Singapore: Springer Nature Singapore, 2023.

[4] Saravanan, T., Saravanakumar, S., Rathinam, G. O. P. A. L., Narayanan, M., Poongothai, T., Patra, P. S. K., & Sengan, S. U. D. H. A. K. A. R. (2022). Malicious attack alleviation using improved time-based dimensional traffic pattern generation in uwsn. Journal of Theoretical and Applied Information Technology, 100(3), 682-689.

[5] Khan, Muhammad Adnan, Wajhe Ul Husnain Abidi, Mohammed A. Al Ghamdi, Sultan H. Almotiri, Shazia Saqib, Tahir Alyas, Khalid Masood Khan, and Nasir Mahmood. "Forecast the influenza pandemic using machine learning." *Computers, Materials and Continua* 66, no. 1 (2020): 331-340.

[6] Singh, Prabh Deep, Rajbir Kaur, Kiran Deep Singh, Gaurav Dhiman, and Mukesh Soni. "Fog-centric IoT based smart healthcare support service for monitoring and controlling an epidemic of Swine Flu virus." *Informatics in Medicine Unlocked* 26 (2021): 100636.

[7] Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita. "Twitter catches the flu: detecting influenza epidemics using Twitter." In *Proceedings of the 2011 Conference on empirical methods in natural language processing*, pp. 1568-1576. 2011.

[8] Saravanakumar, S. (2020). Certain analysis of authentic user behavioral and opinion pattern mining using classification techniques. *Solid State Technology*, *63*(6), 9220-9234.

[9] Kolli, Srinivas, Ahmed J. Obaid, K. Saikumar, and V. Sivakumar Reddy. "An Accurate Swine Flu Prediction and Early Prediction Using Data Mining Technique." In *AI and Blockchain in Healthcare*, pp. 225-237. Singapore: Springer Nature Singapore, 2023.

[10] Saravanakumar, S., & Thangaraj, P. (2019). A computer aided diagnosis system for identifying Alzheimer's from MRI scan using improved Adaboost. Journal of medical systems, 43(3), 76.

[11] Saravanan, T., & Saravanakumar, S. (2022). Enhancing investigations in data migration and security using sequence cover cat and cover particle swarm optimization in the fog paradigm. International Journal of Intelligent Networks, 3, 204-212

[12] Ayachit, Sai Sanjay, Tanmay Kumar, Shriya Deshpande, Nayan Sharma, Kuldeep Chaurasia, and Mayank Dixit. "Predicting h1n1 and seasonal flu: Vaccine cases using ensemble learning approach." In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pp. 172-176. IEEE, 2020.

[13] Kakulapati, V., R. Sai Sandeep, V. Kranthi kumar, and R. Ramanjinailu. "Fuzzy-based predictive analytics for early detection of disease—A machine learning approach." In *ICT Systems and Sustainability: Proceedings of ICT4SD 2020, Volume 1*, pp. 89-99. Springer Singapore, 2021.

[14] Rao, N. Thirupathi. "Prediction of Swine Flu using a Hybrid Voting Algorithm." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12, no. 10 (2021): 1169-1177.

[15] Prabha, I. Surya, and M. Sriram. "Estimation and Prediction of Swine Flu Information using Speech Based Chatbot Model." *International Journal of Intelligent Systems and Applications in Engineering* 12, no. 7s (2024): 298-308.

[16] Inampudi, Srividya, Greshma Johnson, Jay Jhaveri, S. Niranjan, Kuldeep Chaurasia, and Mayank Dixit. "Machine Learning Based Prediction of H1N1 and Seasonal Flu Vaccination." In *Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I 10*, pp. 139-150. Springer Singapore, 2021.

[17] Thangavel, S., & Selvaraj, S. (2023). Machine Learning Model and Cuckoo Search in a modular system to identify Alzheimer's disease from MRI scan images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 11(5), 1753-1761

[18] Jang, Beakcheol, Inhwan Kim, and Jong Wook Kim. "Effective training data extraction method to improve influenza outbreak prediction from online news articles: deep learning model study." *JMIR Medical Informatics* 9, no. 5 (2021): e23305.

[19] Chaurasia, Kuldeep, and Mayank Dixit. "Machine learning based prediction of h1n1 and seasonal flu vaccination." In *Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part I*, vol. 1367, p. 139. Springer Nature, 2021.

[20] Amin, Samina, M. Irfan Uddin, M. Ali Zeb, Ala Abdulsalam Alarood, Marwan Mahmoud, and Monagi H. Alkinani. "Detecting dengue/flu infections based on tweets using LSTM and word embedding." *IEEE access* 8 (2020): 189054-189068.