Predicting Antiviral Compounds for Avian Influenza A/H9N2 Using Logistic Regression with RBF Kernel

1st Siti Amiroch*
Department of Mathematics
Faculty of Mathematics and Natural Sciences
Universitas Islam Darul 'Ulum
Lamongan, Indonesia
siti.amiroch@unisda.ac.id

2nd Mohammad Jamhuri
Department of Mathematics
Faculty of Science and Data Analytics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
Department of Mathematics
Faculty of Science and Technology
UIN Maulana Malik Ibrahim
Malang, Indonesia

3rd Mohammad Isa Irawan
Department of Mathematics
Faculty of Science and Data Analytics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia

4th Imam Mukhlash
Department of Mathematics
Faculty of Science and Data Analytics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia

5th Chairul Anwar Nidom
Professor Nidom Foundation
Surabaya, Indonesia
Faculty of Veterinary Medicine
Universitas Airlangga
Surabaya, Indonesia

***Abstract*—Avian Influenza A/H9N2 is a significant threat to the global poultry industry and presents occasional but severe health risks to humans. Given the potential ramifications of an outbreak, the swift and accurate identification of effective antiviral compounds becomes crucial. Traditional methods employed for predicting the efficacy of these compounds often encounter challenges, particularly in maintaining a balance between accuracy and efficiency. Recognizing these limitations, our study introduces an innovative predictive approach. We leverage the combined strengths of Radial Basis Function (RBF) networks and Logistic Regression. This methodology transforms compound features using the RBF network. The changed features are then fed into a Logistic Regression model to make predictions regarding efficacy. Initial findings from our research indicate a remarkable enhancement in prediction accuracy and precision compared to prevalent methods. Furthermore, our study provides a potentially transformative tool for antiviral compound prediction and establishes a precedent, emphasizing the profound potential of hybrid modeling techniques in advancing biomedical research.**

***Index Terms*—Avian Influenza A/H9N2, Hybrid machine learning models, Log-RBF methodology, Antiviral compound prediction, Drug repurposing.**

I. Introduction

Avian influenza A/H9N2, commonly called bird flu, is a significant concern for the poultry industry and public health [1]. It is highly contagious among birds and can occasionally be transmitted to humans, leading to a need for effective countermeasures to address the potential global implications of this virus.

Identifying effective antiviral compounds is crucial for global health. This is especially important when vaccines may not work against viruses like A/H9N2 [2]. Our research highlights the significance of this approach in reducing the impact of diseases and improving global public health security [3].

Historically, we relied on tangible experimental methodologies for drug discovery and development [4], [5]. However, with the advent and progression of computational techniques, our inclination has shifted toward in-silico methods. Virtual screening, a distinguished computational methodology, facilitates the identification of molecular structures likely to bind to specific drug targets [6], [7]. Despite the merits of these techniques, we acknowledge their inherent limitations. On the other hand, machine learning can efficiently predict binding affinity based on patterns in the data without explicitly modeling the physical interactions [8], [9]. Algorithms such as the support vector machine (SVM), random forests, gradient boosting, multilayer perceptron (MLP), and logistic regression (LR) often face challenges when deciphering the multifaceted, non-linear dynamics intrinsic to biological data [10].

We propose the Log-RBF method, an innovative approach that combines the precision of Logistic Regression [11] with the robust capabilities of the Kernel Radial Basis Function Multiquadratic [12]. This hybrid model leverages the strengths of both techniques and offers potential solutions to the challenges of existing predictive methodologies. We aim to identify active compounds effective against avian influenza A/H9N2 using the Log-RBF method. In this context, our approach represents a paradigm shift in the search for antiviral compounds. It leverages the extensive chemical data available in the public domain [13] and adheres to the principles of drug

*Corresponding author: Siti Amiroch. siti.amiroch@unisda.ac.id

979-8-3503-6935-9/23/$31.00 ©2023 IEEE

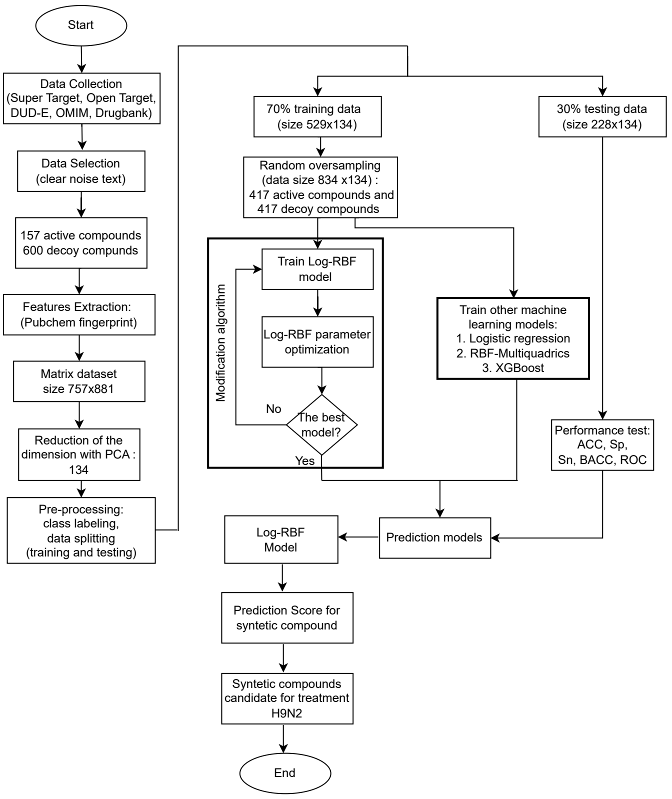
Start
Data Collection (Super Target, Open Target, DUD-E, OMIM, Drugbank)
Data Selection (clear noise text)
157 active compounds 600 decoy compunds
Features Extraction: (Pubchem fingerprint)
Matrix dataset size 757x881
Reduction of the dimension with PCA : 134
Pre-processing: class labeling, data splitting (training and testing)
70% training data (size 529x134)
30% testing data (size 228x134)
Random oversampling (data size 834 x134) : 417 active compounds and 417 decoy compounds
Modification algorithm
Train Log-RBF model
Log-RBF parameter optimization
The best model?
No
Yes
Train other machine learning models: 1. Logistic regression 2. RBF-Multiquadrics 3. XGBoost
Performance test: ACC, Sp, Sn, BACC, ROC
Prediction models
Log-RBF Model
Prediction Score for syntetic compound
Syntetic compounds candidate for treatment H9N2
End

Fig. 1: The research methodology framework

repurposing [14].

This new approach accelerates the identification of antiviral agents, reducing traditional drug discovery time and cost, and enhances the understanding and application of machine learning in complex biological systems [15]. Our research provides an efficient and cost-effective method for identifying potential antiviral compounds effective against A/H9N2, setting a precedent for responding to future viral outbreaks.

II. Material and Method

Figure 1 presents the research methodology framework we used to predict candidate antiviral compounds employing Logistic Regression, RBF-Multiquadrics, and the XGBoost method. Logistic Regression and XGBoost have been applied previously, and their results can be seen in our work [9]. In this study, we primarily focus on the RBF-Multiquadrics method, using the other methods for comparative purposes.

A. Collection and Selection Data

We obtained data on the H9N2 virus target protein from online databases, including the protein data bank (PDB) accessed on April 1, 2021, and The European Bioinformatics Institute (EBI) accessed on April 1, 2021. From these datasets, we identified five significant proteins of the H9N2 virus: protein basic polymerase2 (PB2), protein basic polymerase1 (PB1), protein polymerase acid (PA), hemagglutinin (HA), and neuraminidase (NS) [16]. These proteins are confirmed targets for active compounds inhibiting avian influenza A/H9N2 virus [17]. Given the established significance of these five proteins as targets of the H9N2 virus [16], we focus on identifying key compounds related to these proteins.

B. Data Preparation

We have a set of crucial steps to develop and optimize data, each tailored to ensure high-quality data and effective modeling. Before training our classification model, we meticulously performed these procedures.

- **Data Selection:** The dataset was made reliable and consistent by thoroughly identifying and clearing any noise text after collection [18].
- **Data Composition:** Our finalized dataset comprised 157 active and 600 decoy compounds. This selection aims to provide a balanced representation, which is crucial for effective modeling.
- **Feature Extraction:** We then extracted critical features from the dataset using the Pubchem fingerprint method [19]. This step transformed the raw data into a format suitable for machine learning algorithms, focusing on the most salient features.
- **Data Matrix:** Post-feature extraction, our dataset assumed a matrix form of size 757×881, representing compounds against their extracted features.
- **Dimensionality Reduction:** Given the vastness of the feature set, we employed Principal Component Analysis (PCA) to reduce the dimensionality of our dataset [20]. This step retained the most critical information while reducing the data size and reducing dimensionality 134.

In the concluding phase, we labeled the data into respective classes, ensuring clear demarcation between active and decoy compounds [21]. Subsequently, the dataset was divided into training and test sets to evaluate previously unseen data robustly. We meticulously prepared and processed our dataset to ensure strong model performance and generalization.

C. Prediction using Logistic Regression-RBF

We use the Logistic Regression-Radial Basis Function (RBF) approach to improve prediction and classification accuracy, which combines the strengths of logistic regression and the RBF network. This approach uses a non-linear transformation of the input space for function approximation.

- **Data Representation**: Given a dataset, the input features and target outputs are represented in Equation (1), defining the matrix representations of input and output data.

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,d} \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \tag{1}$$

- **Feedforward Process**: The input data is passed through the RBF network, transforming the feature space as per Equation (2).

$$z_i = w_0 + \sum_{j=1}^{n} w_j \phi(\mathbf{x}_i, \mathbf{c}_j) \tag{2}$$

where ϕ is the multiquadrics function, and $\mathbf{c}_j$ represents the j-th center of the RBF [22].

- **Activation using Logistic Regression**: The transformed features are fed into the logistic regression model, resulting in the output described by Equation (3).

$$s_i = \frac{1}{1+\exp(-z_i)} \tag{3}$$

- **Backpropagation**: The backpropagation process adjusts the weights based on the error calculated between predicted and actual outputs, as detailed in Equation (4).

$$w_j^{(t+1)} = w_j^{(t)} - \alpha \frac{\partial L}{\partial w_j}\bigg|^{(t)} \tag{4}$$

where L is the loss function, α is the learning rate, and t denotes the iteration number.

- **Regularization**: Regularization is introduced to prevent overfitting, as formulated in Equation (5), by adding a regularization term to the loss function.

$$L = \sum_{i=1}^{m} l(y_i, s_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{5}$$

where λ is the regularization parameter.

- **Prediction**: The final prediction step uses the trained model to predict output for new inputs, following the formula in Equation (6).

$$y_i = \frac{1}{1+\exp(-z_i)} \tag{6}$$

Through this method, the Logistic Regression-RBF offers a rigorous and comprehensive approach to classification and prediction, promising enhanced accuracy and reliability.

D. Evaluation Criteria and Measuring Tools

Our classification model's evaluation employs various standard and advanced metrics.

- **Confusion Matrix**: The confusion matrix is a fundamental tool for assessing a model's predictions against actual outcomes. It comprises True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), each offering insights into specific aspects of the model's predictive capabilities.
- **Accuracy (ACC)**: Equation (7) measures the model's overall accuracy, considering both positive and negative correct predictions.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

- **Sensitivity (Sn)**: Sensitivity, or True Positive Rate, is calculated per Equation (8), indicating the model's ability to identify positive instances correctly.

$$S_n = \frac{TP}{TP+FN} \tag{8}$$

- **Specificity (Sp)**: Specificity measures the accuracy in classifying negative instances, as shown in Equation (9).

$$S_p = \frac{TN}{TN+FP} \tag{9}$$

- **AUC/ROC**: The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate against the False Positive Rate across different thresholds. The area under this curve (AUC) is a scalar representation of the model's discriminative power. An AUC closer to 1 indicates the superior model.
- **Balanced Accuracy (BACC)**: Balanced Accuracy, computed using Equation (10), provides an accuracy measure for potential class imbalances.

$$BACC = \frac{S_n+S_p}{2} \tag{10}$$

These metrics ensure a comprehensive and multifaceted evaluation of our classification model.

III. Results

Previous research [9] utilized various machine learning algorithms, including Logistic Regression, k-Nearest Neighbors, Support Vector Machine, Multilayer Perceptron, Random Forest, Gradient Boosting, and XGBoost, for virtual screening. Synthetic active compounds were used to identify potential antiviral candidates against H9N2. The current study introduces a novel method, Log-RBF, and compares its test parameters with those of machine learning techniques such as XGBoost, Logistic Regression, and RBF-Multiquadratic.

A. Model Building and Validation

Log-RBF is a novel method that enhances Logistic Regression by integrating it with the Radial Basis Function kernel. This hybrid model transforms input data into a high-dimensional feature space, overcoming the linear limitations of Logistic Regression and improving its performance with non-linear datasets.

Table I shows the ideal training parameters for the Log-RBF model after extensive testing.

TABLE I: Experimental Results for Parameter Selection

α	Iterations	λ	Accuracy	CT (seconds)
1×10^{-4}	100	0.3	0.9079	29.34
	1000	0.3	0.9211	301.88
	3000	0.3	0.9123	1026.49
	5000	0.3	0.9298	1761.53
	5000	0.1	0.9386	1477.70
	7000	0.1	0.9386	2092.25
1×10^{-3}	8000	0.05	0.9518	2411.87

Where α, λ, and CT in Table I represent the learning rate, regularization, and computational times, respectively. The results in Table I indicate that a parameter with high accuracy was selected for training the data, as shown in Table II.

TABLE II: Parameters Used for Training Process

Parameter	Description	Value
α	Learning rate	0.001
λ	Constant for regularization	0.05
Epoch	Number of iterations used	200

The resulting graph from the selected parameters in Table II is shown in Figure 2.

Figure 2 demonstrates that the accuracy of the training data is not significantly different from the accuracy of the testing data. At epoch 200, the accuracy of the testing data converges with that of the training data. However, from epoch 200 onwards, the accuracy of the training data exceeds that

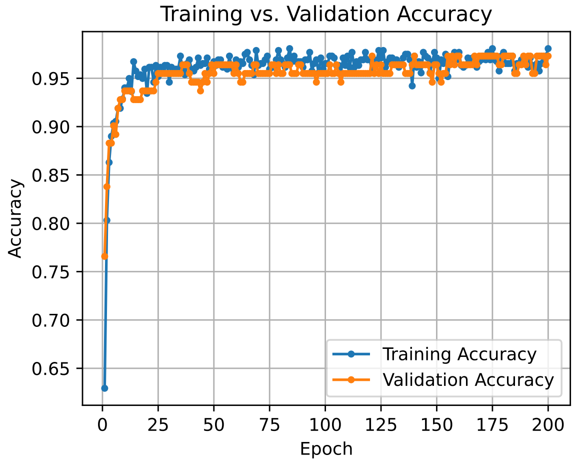

Fig. 2: Visualization of training and validation data accuracy with $\alpha = 0.001$, $\lambda = 0.05$, and epoch=200

of the testing data. The accuracy of the testing data tends to be more stable across each epoch, increasing above 0.95 as the number of epochs rises. Plotting a graph to compare the accuracy of the training and testing processes reveals that the gap between the two is not significant. This indicates that the model we developed does not exhibit symptoms of overfitting, which is characterized by decreasing error during training (with larger epochs) but increasing error during testing.

B. Performance Evaluation

The model's performance was evaluated using 30% of the compound data. This proportion was determined based on experiments conducted with various data proportions, as shown in Table III. The results from the optimal proportion were then applied in this study.

TABLE III: Testing results of training and testing data proportions in the Log-RBF method

Proportion (Log-RBF)	ACC	Sn	Sp	ROC	BACC
80% : 20%	0.8224	0.7812	0.9212	0.8512	0.8512
75% : 25%	0.9478	0.8792	0.9573	0.9183	0.9183
70% : 30%	0.9518	0.8696	0.9725	0.9210	0.9210
65% : 35%	0.9286	0.8892	0.9413	0.9153	0.9153
60% : 40%	0.9247	0.8893	0.9542	0.9218	0.9218

Based on the test results in Table III, the best proportion for training and testing data using the Log-RBF method was found to be 70% : 30%. Consequently, this proportion was used as the benchmark for training and testing the data.

The following compares the performance of Logistic Regression, Radial Basis Function (using Multiquadratic kernel), Log-RBF, and XGBoost models using training and testing data.

TABLE IV: Comparison of model performance for test data results

Classification Model	ACC	Sn	Sp	ROC	BACC
LR	0.9474	0.8666	0.9672	0.9169	0.9169
RBF-Multiquadratic	0.8465	0.8394	1	0.9197	0.9197
Log-RBF	0.9518	0.8696	0.9725	0.9210	0.9210
XGBoost	0.9649	0.8222	1	0.9111	0.9111

In Table IV, Log-RBF outperforms Logistic Regression and RBF-Multiquadratic in terms of accuracy, sensitivity, ROC, and BACC, but is slightly outperformed by RBF-Multiquadratic and XGBoost in terms of specificity. Specificity is the true negative rate. While a higher specificity value indicates better prediction for the negative class, in this context, where both positive and negative classes are predicted, a slightly lower specificity value for Log-RBF compared to RBF-Multiquadratic and XGBoost is not significantly detrimental. The Log-RBF model in Table IV shows slightly lower accuracy than XGBoost, but outperforms it in terms of sensitivity, ROC score, and BACC. Overall, as a modification between Logistic Regression and RBF-Multiquadratic, Log-RBF demonstrates a significant performance improvement over both Logistic Regression and RBF-Multiquadratic.

C. Prediction Results of Synthetic Compound Candidates

The Log-RBF model predicted synthetic compound data, totaling 157, and verified herbal compound data, totaling 845. From the prediction results on synthetic compounds, with a threshold of 0.5, 151 compounds were identified. A more specific threshold of 0.992 resulted in 124 compounds. Table V below lists the top 30 synthetic compounds as predicted by the Log-RBF model.

TABLE V: List of synthetic compounds with the highest Log-RBF prediction results

No.	Compound Name	C ID	Score
1	Ligan C	6442269	0.999884603
2	Ligan D	6912404	0.999794484
3	Laninamivir O	9847629	0.999785829
4	Pyrrolidine D	5329293	0.999710169
5	Zanamivir	20112027	0.999668209
6	Zanamivir	60855	0.999668209
7	4-Amino-N	445533	0.999657643
8	Deoxysialic	65309	0.999652969
9	Pyridine D40	5278296	0.999573566
10	Pyrrolidine D34	5329301	0.99956914
11	Cyclopentane D16g	5329067	0.99950569
12	2,4-deoxy 4G	5288452	0.999429872
13	AC1NQT9J	5278609	0.999293296
14	AC1NQT9P	5278611	0.999293296
15	AC1NQT9V	5278613	0.999293296
16	AC1NQTAA	5278618	0.999293296
17	Cyclopentane D16f	5329066	0.999293296
18	AC1NQT9Y	5278614	0.999166495
19	Benzoic Acid deriv. 6b	506044	0.999137395
20	Benzoic Acid deriv. 149	506095	0.999098228
21	Pyrrolidine deriv. 24	5329292	0.999045962
22	BANA 113	446323	0.998998429
23	4-acetamido A	446367	0.998966553
24	AC1NQT8M	5278598	0.998931435
25	AC1NQT8P	5278599	0.998931435
26	AC1NQT8Y	5278602	0.998931435
27	AC1NQT91	5278603	0.998931435
28	AIDS292405	5278607	0.998931435
29	AC1NQTA4	5278616	0.998919372
30	AC1NQTA7	5278617	0.998912794

”Ligan C” and ”Ligan D” are abbreviations used for ligands. Table VI compares the prediction results of the XGBoost model with those of the Log-RBF model for the same set of compounds, including the top 30 compounds from the XGBoost prediction results.

As shown in Table V, and further in Table VI, the prediction results vary. For example, for the compound benzoic acid

TABLE VI: Comparison of compound rankings and scores between XGB and log-RBF methods.

Compound Name	Pubchem Id	Rank XGB	Score XGB	Rank Log-RBF	Score Log-RBF
AC1NQT9A	5278606	1	0.9998	43	0.9985
AC1NQT8D	5278595	2	0.9997	48	0.9983
AC1NQT8G	5278596	3	0.9997	49	0.9983
AIDS292422	5278601	4	0.9997	50	0.9983
AC1NQT9G	5278608	5	0.9997	56	0.9980
AC1NQT8M	5278598	6	0.9997	24	0.9989
AC1NQT8P	5278599	7	0.9997	25	0.9989
AC1NQT8Y	5278602	8	0.9997	26	0.9989
AC1NQT91	5278603	9	0.9997	27	0.9989
AIDS292405	5278607	10	0.9997	28	0.9989
2,4-DEOXY-4	5288452	11	0.9997	12	0.9994
AC1NQT7P	5278587	12	0.9997	32	0.9988
AC1NQT7S	5278588	13	0.9997	33	0.9988
AC1NQT7V	5278589	14	0.9997	34	0.9988
AC1NQT7Y	5278590	15	0.9997	35	0.9988
AIDS292384	5278586	16	0.9997	36	0.9988
AC1NQT77	5278581	17	0.9996	51	0.9981
AC1NQT7A	5278582	18	0.9996	52	0.9981
AC1NQT7G	5278584	19	0.9996	53	0.9981
AC1NQT7J	5278585	20	0.9996	54	0.9981
AC1NQT9J	5278609	21	0.9996	13	0.9992
AC1NQT9P	5278611	22	0.9996	14	0.9992
AC1NQT9V	5278613	23	0.9996	15	0.9992
AC1NQTAA	5278618	24	0.9996	16	0.9992
Cyclo. P16f	5329066	25	0.9996	17	0.9992
AC1NQT8A	5278594	26	0.9995	41	0.9987
AC1NQTA4	5278616	27	0.9995	29	0.9989
Pyrrolidine	5329298	28	0.9995	91	0.9911
Acetylamino	446326	29	0.9993	95	0.9894
Benzoic AI7	5275967	30	0.9993	76	0.9952

inhibitor 7 with pubchem_ID 5275967, XGBoost ranks it 30th, while Log-RBF ranks it 76th. The ranking difference is significant, but the prediction score difference is only 0.004092446. A random comparison of a few compounds, as in Table VII, shows that the prediction results of XGBoost and Log-RBF on some potential compounds are very close.

Table VII illustrates that the differences between XGBoost and Log-RBF predictions on some potential compounds are minimal

IV. Discussion

When dealing with non-linear datasets, logistic regression faces various challenges. The Log-RBF method acknowledges these challenges and proposes a solution. By merging the Radial Basis Function (RBF) kernel with logistic regression, we have addressed the limitations of each method while capitalizing on their strengths. This synergy is crucial to enhance the prediction of antiviral compounds for Avian Influenza A/H9N2 in our study [23].

Our results, as evidenced in Tables I, II and Figure 2, demonstrate the efficacy of this approach. The convergence in accuracy between training and testing datasets around epoch 200, as illustrated in Figure 2, is particularly noteworthy. It suggests the model generalizes well to new data without overfitting or underfitting.

The meticulous optimization of the Log-RBF model's parameters reflects a tailored approach to this dataset. The consistent performance across epochs, particularly after the 200th, underscores the model's robustness see Table II. This stability in the testing data's accuracy, particularly its maintenance above 0.95, suggests effective data pattern capture.

However, the minimal divergence in accuracy post the 200th epoch reminds us of the dynamic nature of machine learning models [24]. It underscores the importance of continuous monitoring and potential recalibration, especially in practical applications.

The Log-RBF method represents a significant step in addressing the real-life challenge of rapid antiviral compound discovery. Its efficiency in identifying potential compounds is crucial in public health contexts, particularly during outbreaks [25]. This approach could significantly reduce the time and resources needed in the initial stages of pharmaceutical development, thereby accelerating response times in public health emergencies.

Furthermore, our study contributes to the broader understanding of machine learning in biomedicine. By applying the Log-RBF method to a complex biological problem, we demonstrate its practicality and effectiveness in a real-world context [26]. This advancement not only paves the way for future research but also opens doors to myriad applications beyond the scope of our current investigation.

The Log-RBF method improves binary classification for non-linear datasets and has potential in practical applications such as healthcare and pharmaceuticals. It can be refined and tested on larger datasets with advanced techniques like deep learning for better performance.

V. Conclusion

Our study has shown that the Log-RBF method is a reliable and promising alternative for predicting effective antiviral compounds against Avian Influenza A/H9N2. It outperforms traditional methods like Logistic Regression and RBF-Multiquadratic regarding accuracy, sensitivity, ROC score, and BACC. Despite differences in prediction rankings, the Log-RBF method also achieved similar prediction scores as the XGBoost model.

This study significantly contributes to the ongoing efforts to combat Avian Influenza A/H9N2. The Log-RBF method identified 124 potential antiviral compounds with a high threshold of 0.992, displaying strong binding affinities and promising pharmacological profiles. These compounds, which require further in vitro and in vivo validation, could serve as vital agents in the battle against the H9N2 virus.

Our work demonstrates the value of machine learning in drug discovery. The Log-RBF method models non-linear feature spaces and provides interpretable results, making it a powerful tool for addressing biological data challenges. However, this begins a long journey toward effective antiviral solutions. Future research should refine the Log-RBF methodology, incorporate more diverse chemical entities, and collaborate with experimental researchers for compound validation.

The fight against Avian Influenza A/H9N2 continues, but with advancements like the Log-RBF method, we are better equipped to tackle this challenge and develop effective therapeutic solutions.

TABLE VII: Comparison of some potential compounds at random

No	Compound Name	C_ID	Rank XGBoost	Score XGBoost	Rank Log-RBF	Score Log-RBF
1	Oseltamivir Carboxylate	449381	62	0.9984663	86	0.993046299
2	Benzoic acid derive 130	506090	67	0.9982233	87	0.992967571
3	Benzoic acid inhibitor 6	1708	72	0.9980004	42	0.998629097
4	AC1NQT84	5278592	84	0.9970549	68	0.99619591
5	Laninamivir Octanoate	9847629	103	0.9958969	3	0.999785829

REFERENCES

[1] Y. Guan, K. F. Shortridge, S. Krauss, and R. G. Webster, "Molecular characterization of h9n2 influenza viruses: were they the donors of the "internal" genes of h5n1 viruses in hong kong?," *Proceedings of the National Academy of Sciences*, vol. 96, no. 16, pp. 9363–9367, 1999.

[2] J. Peiris, Y. Guan, D. Markwell, P. Ghose, R. Webster, and K. Shortridge, "Cocirculation of avian h9n2 and contemporary "human" h3n2 influenza a viruses in pigs in southeastern china: potential for genetic reassortment?," *Journal of virology*, vol. 75, no. 20, pp. 9679–9686, 2001.

[3] S. Amiroch, M. I. Irawan, I. Mukhlash, A. N. M. Ansori, and C. A. Nidom, "Identification of the spread of the influenza virus type a/h9n2 in indonesia using the neighbor-joining algorithm with felsenstein models," *Journal of Hunan University Natural Sciences*, vol. 48, no. 5, 2021.

[4] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, "Machine learning in chemoinformatics and drug discovery," *Drug discovery today*, vol. 23, no. 8, pp. 1538–1546, 2018.

[5] P. Carracedo-Reboredo, J. Liñares-Blanco, N. Rodríguez-Fernández, F. Cedrón, F. J. Novoa, A. Carballal, V. Maojo, A. Pazos, and C. Fernandez-Lozano, "A review on machine learning approaches and trends in drug discovery," *Computational and structural biotechnology journal*, vol. 19, pp. 4538–4558, 2021.

[6] B. Rizzuti and F. Grande, "Virtual screening in drug discovery: A precious tool for a still-demanding challenge," in *Protein Homeostasis Diseases*, pp. 309–327, Elsevier, 2020.

[7] K. A. Carpenter and X. Huang, "Machine learning-based virtual screening and its applications to alzheimer's drug discovery: a review," *Current pharmaceutical design*, vol. 24, no. 28, pp. 3347–3358, 2018.

[8] M. Kontoyianni, "Docking and virtual screening in drug discovery," *Proteomics for drug discovery: Methods and protocols*, pp. 255–266, 2017.

[9] S. Amiroch, M. I. Irawan, I. Mukhlash, M. H. Z. Al Faroby, and C. A. Nidom, "Machine learning for the prediction of antiviral compounds targeting avian influenza a/h9n2 viral proteins," *Symmetry*, vol. 14, no. 6, p. 1114, 2022.

[10] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20170387, 2018.

[11] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398. John Wiley & Sons, 2013.

[12] D. Broomhead and D. Lowe, "Multivariable functional interpolation and adaptive networks, complex systems, vol. 2," 1988.

[13] N. Brown, J. Cambruzzi, P. J. Cox, M. Davies, J. Dunbar, D. Plumbley, M. A. Sellwood, A. Sim, B. I. Williams-Jones, M. Zwierzyna, *et al.*, "Big data in drug discovery," *Progress in medicinal chemistry*, vol. 57, pp. 277–356, 2018.

[14] S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, *et al.*, "Drug repurposing: progress, challenges and recommendations," *Nature reviews Drug discovery*, vol. 18, no. 1, pp. 41–58, 2019.

[15] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, and S. Drăghici, "Machine learning and its applications to biology," *PLoS computational biology*, vol. 3, no. 6, p. e116, 2007.

[16] A. Samir, A. Adel, A. Arafa, H. Sultan, and H. A. Hussein Ahmed, "Molecular pathogenic and host range determinants of reassortant egyptian low pathogenic avian influenza h9n2 viruses from backyard chicken," *International Journal of Veterinary Science and Medicine*, vol. 7, no. 1, pp. 10–19, 2019.

[17] X. Sun, J. Belser, and T. Maines, "Adaptation of h9n2 influenza viruses to mammalian hosts: a review of molecular markers. viruses. 2020; 12 (5): 541."

[18] C. Pete, C. Julian, K. Randy, K. Thomas, R. Thomas, S. Colin, and R. Wirth, "Crisp-dm 1.0—step-by-step data mining guide," *Cris. Consort*, p. 76, 2000.

[19] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, "Pubchem: integrated platform of small molecules and biological activities," in *Annual reports in computational chemistry*, vol. 4, pp. 217–241, Elsevier, 2008.

[20] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[21] S. Raschka and V. Mirjalili, *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019.

[22] J. Ghosh and A. Nag, "An overview of radial basis function networks," *Radial basis function networks 2: new advances in design*, pp. 1–36, 2001.

[23] E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1178–1192, 2022.

[24] W. Gilpin, Y. Huang, and D. B. Forger, "Learning dynamics from large biological data sets: machine learning meets systems biology," *Current Opinion in Systems Biology*, vol. 22, pp. 1–7, 2020.

[25] A. von Delft, M. D. Hall, A. D. Kwong, L. A. Purcell, K. S. Saikatendu, U. Schmitz, J. A. Tallarico, and A. A. Lee, "Accelerating antiviral drug discovery: lessons from covid-19," *Nature Reviews Drug Discovery*, pp. 1–19, 2023.

[26] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, *et al.*, "Applications of machine learning in drug discovery and development," *Nature reviews Drug discovery*, vol. 18, no. 6, pp. 463–477, 2019.