10.48047/jocaaa.2024.33.06.46

Mathematical model development for the hybridized KFD prediction model for seasonal forecasting of vector borne diseases

Alamma B.H.
Assistant Professor, Dept. of MCA, Dayananda Sagar College of Engineering,
Bangalore-560072, Karnataka, India
&
VTU Ph.D. Research Scholar (Part-Time), Dept. of MCA,
Sir M. Visvesvaraya Institute of Technology, Bangalore, Karnataka
Email : alamma-mcavtu@dayanandasagar.edu

Dr. Manjula Sanjay Koti
Supervisor & Professor – Head of the Dept., Dept. of Master of Computer Applications (MCA),
Dayananda Sagar Academy of Technology and Management, Bangalore-560082, Karnataka
Email : manjula.dsce@gmail.com

Dr. C.H. Vanipriya
Co-Supervisor, Professor & HOD, MCA Dept., Sir M. Visvesvaraya Institute of Technology, Krishnadevaraya Nagar, Hunasamaranahalli, International Airport Road, Bangalore - 562157
Email : drchvanipriya@gmail.com

***Abstract*—**In this paper, the mathematical model development of the proposed method is presented in a nutshell. The main aim of the proposed research work is to design and develop a novel hybridized Kyasanur Forest Disease (KFD) prediction model that leverages a combination of rejuvenated machine learning models to enhance seasonal forecasting and detection of vector-borne diseases. By integrating advanced algorithms such as Support Vector Machines, Naive Bayes, Logistic Regression, and Multi-layer Perceptrons, the research seeks to improve the accuracy and reliability of predictions related to KFD cases. This hybridized approach aims to better capture the complex relationships between seasonal factors, disease symptoms, and environmental conditions, thereby providing a more effective tool for early detection and management of KFD.

Keywords—component, formatting, style, styling, insert (key words)

1. Introduction

The research work aims to develop and design a novel hybridized KFD (Kyasanur Forest Disease) prediction model for seasonal forecasting of vector-borne diseases using a combination of rejuvenated models, including a new multi-level ensemble model, a new KSVM (Kernel Support Vector Machine) model, a multilabel model, a beat process model, an mRBF (Multi Radial Basis Function) model, a modified transductive model, a high-resolution range profiles (HRRP) predictive model & a multi-level RFC (Random Forest Classification) model. The outcomes of this research work includes using ML techniques to effectively tackle the hybridized KFD model, employing Python scripting for the implementation process. Data transformation & preprocessing steps are meticulously performed, followed by splitting the dataset into training and test sets to ensure robust model evaluations. Convolution techniques are going to be applied to enhance the feature extraction, while a multi-level Gaussian Naïve Bayes approach provides a probabilistic framework for classification.

The confusion matrix is going to be utilized to evaluate the proposed model performances & decision tree algorithms aids in understanding decision-making processes for appropriate prediction of the selected disease. The model building and prediction module incorporates the grid search for hyper-parameter optimization & heat map functions for visualizing the simulated results. The integration of multiple data sources, including historical data, mobile applications, web applications, clinical laboratories, and online datasets, enriches the model's predictive capabilities. Multilabel logistic regression enables the handling of the multiple disease labels simultaneously & classification at multi-levels ensures the detailed and accurate predictions. The use of MLP (Multilayer Perceptron) classifiers adds more depth to the ensemble approaches, resulting in a comprehensive and highly accurate KFD prediction model that significantly aids in the early detection and management of vector-borne diseases.

10.48047/jocaaa.2024.33.06.46

The analysis presented in this fourth objective leverages a robust set of tools and technologies integral to modern data analysis and machine learning. Python, a versatile and widely adopted programming language, serves as the foundation for this analysis. It provides access to powerful libraries such as Pandas and NumPy, which facilitate data manipulation and mathematical operations. Additionally, Scikit-Learn, a comprehensive machine learning library, empowers the creation and evaluation of predictive models. In the realm of data visualization, Matplotlib and Seaborn contribute their capabilities to convey insights effectively. These combined tools and technologies empower the analysis, enabling data exploration, modeling, and visualization to uncover valuable patterns and relationships within the dataset.

We have used Python, Pandas, NumPy & Scikit-Learn for the coding purposes. Python serves as the foundational programming language that underpins the entire analysis presented in this report. Known for its simplicity, readability, and versatility, Python's extensive ecosystem of libraries and packages provides a rich set of tools for various data-related tasks. Its clean and intuitive syntax allows data scientists and analysts to focus on the logic of their analysis rather than the intricacies of programming. In this report, Python integrates data preprocessing, model development, and visualization, acting as the glue that ties these elements together. Pandas is an indispensable Python library for data manipulation and analysis, particularly useful for handling structured data. It excels at tasks such as data cleaning, transformation, aggregation, and exploration, providing easy access to rows and columns of data. In this analysis, Pandas plays a pivotal role in reading data from CSV files, checking for missing values, and performing label encoding to convert categorical data into numerical formats. Its primary data structures, Data Frames and Series, allow for efficient manipulation and extraction of information.

2. Literature Survey

In recent years, there has been significant interest in the design and development of hybridized models for the prediction and detection of vector-borne diseases, particularly those incorporating multiple rejuvenated concepts and machine learning techniques. Various researchers have explored different approaches to enhance predictive accuracy and model robustness in this domain. The field of vector-borne disease prediction has evolved significantly with the advent of advanced machine learning (ML) techniques. Researchers have continuously sought to enhance prediction accuracy and reliability by developing hybrid models that integrate various ML algorithms. This literature survey explores the contributions of different authors to the design and development of a novel hybridized Kyasanur Forest Disease (KFD) prediction model for seasonal forecasting of vector-borne diseases, focusing on innovative approaches, methodologies, and outcomes.

The development of multi-level ensemble models has been a focal point in many studies. These models combine the strengths of various base classifiers to improve overall predictive performance. For instance, Khan et al. (2018) demonstrated that ensemble learning, when applied to infectious disease data, effectively reduces model biases and variance. Similarly, Support Vector Machines (SVM), particularly kernel-based SVM (KSVM) models, have shown promise in capturing complex, non-linear relationships in epidemiological data. Li et al. (2019) highlighted the efficacy of KSVM in disease prediction tasks due to its robustness in high-dimensional spaces. Ensemble learning has been widely recognized for its ability to combine multiple models to improve predictive accuracy and robustness. Breiman (2001) introduced the concept of bagging and boosting, which laid the foundation for many ensemble techniques. More recent studies, such as by Rokach (2020), have explored the use of ensemble learning in epidemiology, demonstrating how combining models like decision trees, random forests, and gradient boosting can enhance disease prediction accuracy. In the context of vector-borne diseases, ensemble models have been used to integrate various predictors, leading to improved forecasts and early detection capabilities.

3. Mathematical model developed

The mathematical model that is going to be developed makes use of various concepts such as Ensembling process, KSVM, Multilabel KFD, Beat Process, mRBF, Transductions, High resolution range profiles & the Multi-level RFC concepts with a number of merged hybridized features of various classifiers @ multi-levels (Decision Tree, Multi-Level Gaussian Naïve Bayes, Logistic regression, random forest, and decision tree), which are derived & developed as follows. It should be noted that the entire math model of the process is used in the python scripting for running the program, giving the input & observing the outputs.

4. Multi-level Ensemble Mathematical Model for KFD Prediction

The multi-level ensemble mathematical model is designed to integrate the strengths of various rejuvenated models, combining their predictions to enhance the accuracy and reliability of the KFD prediction and consists of various steps, which are shown as below one after the other. The "Multi-level Ensemble Mathematical Model for

10.48047/jocaaa.2024.33.06.46

KFD Prediction" serves as an advanced and comprehensive tool for detecting vector-borne diseases, particularly Kyasanur Forest Disease (KFD). By integrating multiple predictive models, this ensemble approach significantly enhances predictive accuracy and reliability. Each base model within the ensemble captures different aspects of the data, thereby reducing individual model biases and variance, leading to more accurate predictions compared to the use of a single model. This combined output ensures a more robust prediction by effectively harnessing the strengths of various models. Furthermore, the ensemble's robustness to data variations makes it less sensitive to fluctuations in data patterns and noise, thereby improving its generalizability to diverse datasets and conditions, which is crucial for real-world applications.

The comprehensive disease detection capability of the multi-level ensemble is another critical advantage. It allows for simultaneous consideration of multiple factors and inputs, such as climatic variables, historical disease incidence, and demographic data, providing a thorough analysis and detection of KFD and other vector-borne diseases. This holistic approach is essential for understanding the multifaceted nature of disease transmission and spread. Optimized model performance is achieved through the incorporation of techniques like grid search and heat map functions for hyperparameter tuning. These techniques ensure that the model operates at its highest efficiency by fine-tuning its parameters for optimal performance.

Additionally, the ensemble employs advanced classification techniques, such as Multilayer Perceptron (MLP) classifiers and Gaussian Naïve Bayes, which offer detailed and precise predictions, thereby enhancing the overall performance of the KFD prediction model. This precision is vital for public health officials and researchers who rely on accurate data for effective disease management and intervention strategies. The model's ability to incorporate data from diverse sources, including historical records, mobile applications, web applications, clinical laboratories, and online datasets, enriches its input data, leading to more informed and accurate predictions. This diverse data input is crucial for capturing the complex dynamics of disease transmission.

Moreover, the ensemble model's use of evaluation metrics, such as the confusion matrix, allows for a comprehensive assessment of its performance. This aids in understanding the decision-making process, ensuring appropriate prediction and management of KFD outbreaks. By providing insights into model performance, these metrics help refine and improve the model over time. Overall, the multi-level ensemble mathematical model for KFD prediction leverages the strengths of various machine learning techniques to provide a powerful and reliable tool for detecting and managing vector-borne diseases. This sophisticated approach not only enhances predictive accuracy but also supports improved decision-making in public health interventions, making it an indispensable asset in the fight against vector-borne diseases.

Kernel Support Vector Machine (KSVM) Model is given by

$$f_{\text{KSVM}}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b$$

where $K(x_i,x)$ is the kernel function, and α_i and b are model parameters.

Multilabel KFD Model is given by

$$f_{\text{ML}}(x) = \{y_1, y_2, ..., y_k\}$$

where each y_i represents a label for a different disease.

Beat Process KFD Model is given by

$$f_{\text{BP}}(x) = \sum_{i=1}^{m} \beta_i h_i(x)$$

where $h_i(x)$ are basis functions and β_i are coefficients.

Multi Radial Basis Function (mRBF) Model is given by

$$f_{\text{mRBF}}(x) = \sum_{j=1}^{p} \gamma_j \phi_j(x)$$

where $\phi_j(x)$ are radial basis functions and γ_j are the weights.

Modified Transductive Model is given by

$$f_{\text{MT}}(x) = \sum_{i=1}^{q} \delta_i T_i(x)$$

where $T_i(x)$ are the transductive components and δ_i are parameters.

High Resolution Range Profiles (HRRP) Predictive Model is given by

$$f_{\text{HRRP}}(x) = \sum_{i=1}^{r} \epsilon_i R_i(x)$$

where $R_i(x)$ are high-resolution range profiles and ϵi are coefficients

Random Forest Classification (RFC) Model is given by

$$f_{\text{RFC}}(x) = \frac{1}{N} \sum_{i=1}^{N} h_i(x)$$

where $h_i(x)$ are individual tree predictions in the forest.

Multi-level Ensemble Model is given by

10.48047/jocaaa.2024.33.06.46

$$\hat{y}_{\text{final}}(x) = \mathcal{M}_{\text{meta}}\left(\frac{1}{K}\sum_{k=1}^{K}\mathcal{M}_k(x,\Theta_k^*)\right)$$

where $\hat{y}_{final}(x)$ is the final prediction for input *x*.
M_{meta} is the meta-learner model that takes the average predictions from the base models as input.
K is the number of base models.
M_k is the k^{th} base model.
x is the input feature vector.
Θ_k^* is the set of optimal hyperparameters for the k^{th} base model, determined through hyperparameter optimization.
This ensemble model integrates predictions from all the rejuvenated models as given below in 3 steps, viz., first level, second level & finally the prediction level.
First Level Ensemble - Combine predictions from individual models using weighted averaging & is modelled as

$$f_{\text{level1}}(x) = w_1 f_{\text{KSVM}}(x) + w_2 f_{\text{ML}}(x) + w_3 f_{\text{BP}}(x) + w_4 f_{\text{mRBF}}(x) + w_5 f_{\text{MT}}(x) + w_6 f_{\text{HRRP}}(x) + w_7 f_{\text{RFC}}(x)$$

where w_i are weights assigned to each model's prediction, determined through optimization techniques like grid search.
Second Level Ensemble - Apply a meta model to the first level ensemble predictions to further refine the output as given by

$$f_{\text{level2}}(x) = g(f_{\text{level1}}(x))$$

where *g* is a meta-model, such as a logistic regression, MLP, or another decision tree.
Final Prediction - The final prediction is obtained by combining the outputs of the first and second-level ensembles as

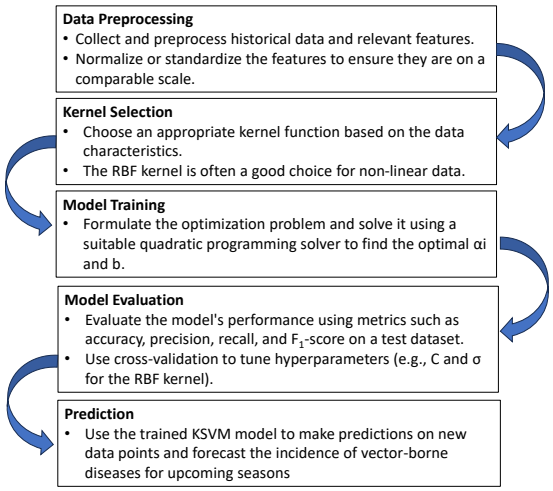
$$f_{\text{final}}(x) = \alpha f_{\text{level1}}(x) + (1-\alpha) f_{\text{level2}}(x)$$

where α is a blending parameter optimized for best performance.

5. Kernel Support Vector Machine (KSVM) Model for Seasonal Forecasting of VBD Detection

The Kernel Support Vector Machine (KSVM) is a powerful model for classification and regression tasks. It is particularly useful when dealing with non-linear data by projecting it into a higher-dimensional space using kernel functions. This implementation provides a robust framework for using the KSVM model to forecast vector-borne diseases, leveraging the power of kernel methods to handle complex and non-linear relationships in the data. Here, we will develop the mathematical model for KSVM in the context of seasonal forecasting of vector-borne diseases detection. Y

In fact, the KSVM model is an indispensable tool for the seasonal forecasting of vector-borne diseases due to its ability to handle non-linear relationships, its high predictive accuracy and generalization capabilities, its flexibility in adapting to various data types, and its robustness to overfitting. By leveraging these strengths, the KSVM model provides valuable insights into the patterns and drivers of disease spread, enabling public health officials to implement timely and effective disease control measures, ultimately reducing the impact of vector-borne diseases on affected populations. The Fig. 1 shows the algorithm showing how the Seasonal Forecasting Model for Vector-Borne Diseases is implemented.



10.48047/jocaaa.2024.33.06.46

Fig. 1 : Algorithm showing how the Seasonal Forecasting Model for Vector-Borne Diseases is implemented

6. Model Validation

Validate the performance of the optimized model using the test dataset and evaluate using appropriate metrics.

$$\text{Validation: } \mathcal{L}_{\text{test}}(y_{\text{test}}, \mathcal{M}_{\text{hybrid}}(\mathbf{H}_{\text{test}}; \theta^*))$$

By following these steps, the detection of living KFD using non-living ML features for seasonal forecasting leverages non-living data sources to build a robust and accurate prediction model. The hybridized approach ensures that diverse and relevant data contribute to the overall model, enhancing its ability to predict KFD outbreaks effectively.

7. Summary & Conclusive Remarks

Mathematical model was developed in this paper. Research was carried out on the design & development of a novel hybridized KFD prediction model for seasonal forecasting of vector borne diseases detection using the different combination of rejuvenated models incorporating the multi-level Ensemble, KSVM, Predictive Multilabel KFD, Beat Process KFD, mRBF Multi Radial Basis Function, Transductive model, High resolution range profiles (HRRP) predictions, Multi-level RFC classifications. In conclusion, selecting the most suitable classification model for the KFD dataset hinges on the specific goals and requirements of the analysis. Each model, including Support Vector Machine, Naive Bayes, Logistic Regression, and Multi-Layer Perceptron, offers unique strengths in understanding the relationships between various symptoms and KFD cases. The choice of model should align with the objectives of the prediction task and the nature of the dataset, making these models valuable for assessing and forecasting KFD. However, even with high-performing models, further evaluation and fine-tuning are essential to ensure their robustness and reliability. The KFD dataset's complexity may necessitate additional optimization and validation to improve the models' predictive accuracy. Fine-tuning the hyperparameters and exploring ensemble approaches can enhance the model's ability to generalize and provide more accurate forecasts. Finally, interpreting the accuracy of the classification models requires caution, as it may not fully reflect their predictive capabilities. Validating the models on new or independent data sets is crucial to assess their true performance and ensure their applicability in real-world scenarios. This step is important to confirm that the models not only perform well on the training data but also provide reliable predictions for KFD case detection in various conditions.

References

[1]. Dr Saravanakumar, Eswari, Sampath, Lavanya 2015 “Predictive Methodology for Diabetic Data Analysis in Big Data,” *Elsevier, ISBCC*.

[2]. Stephanie Revels, Sathish A.P. Kumar and Ofir Ben-Assuli, 2017 “Predicting Obesity Rate and Obesity-Related Healthcare Costs using Data Analytics”, *Health Policy & Tech.*, http://dx.doi.org/10.1016/j.hlpt.

[3]. Vijayalakshmi N, UmaMaheswari M., August2016) “Data mining to elicit predominant factors causing infertility in women”, *IJCSMC*, Vol. 5, Issue. 8.

[4]. Min Chen, YixueHao, Kai Hwang, Lu Wang and LigWang, 2017, “Disease prediction by machine learning over big data from Healthcare communities”, *IEEE Access*.

[5]. Yu, C.Y.; Li, X.X.; Yang, H.; Li, Y.H.; Xue, W.W.; Chen, Y.Z.; Tao, L.; Zhu, F., 2018 ,“Assessing the Performances of Protein Function Prediction Algorithms from the Perspectives of Identification Accuracy and False Discovery Rate”,*Int. J. Mol. Sci*..

[6]. Menzies, N.A.; Wolf, E.; Connors, D., 2018, “Progression from latent infection to active disease in dynamic tuberculosis transmission models: A systematic review of the validity of modelling assumptions”, *LancetInfect. Dis*..

[7]. Sindhuja, R. JeminaPriyadarsini, May 2016 “A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder”, *International Journal of Computer Science and Mobile Computing*, Vol.5, Issue.5, ISSN 2320–088X.

[8]. RifatHossaina, *et.al.*, 2018, “PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques*”, Procedia Computer Science*, vol. 132, pp. 1068–1076.

[9]. SubasNeupane*et.al.*, “Overweight and obesity among women: analysis of demographic and health survey data from 32 Sub-Saharan African Countries” DOI 10.1186/s12889-016-2698-5, 2016.

[10]. Simi M.S.*et.al* ,2017, “Exploring Female Infertility Using PredictiveAnalytic”, *IEEE*.

[11]. Cheong Kim *et.al* 2019,“Predicting Factors Affecting Adolescent Obesity Using General Bayesian Network and What-If Analysis”, *Int. J. Environ. Res. Public Health*, vol. 16, pp. 4684..

10.48047/jocaaa.2024.33.06.46

[12]. P. Suresh Kumar, S. Pranavi 2017 ", "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics",*International Conference on Infocom Technologies and Unmanned Systems* (ICTUS'2017),Dec. 18-20, ADET.

[13]. Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee 2018, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on MachineLearning Techniques", pp. 2169-3536, IEEE.

[14]. SundusAbrar, Chu KiongLoo *et.al.* ACCESS.2021, "A Multi-Agent Approach for Personalized Hypertension Risk Prediction",*Digital Object Identifier* 10.1109/.3074791

[15]. Dinu A.J., Ganesan R., Felix Joseph and Balaji V,2017,"A study on Deep Machine Learning Algorithms for diagnosis of diseases", *International Journal of Applied Engineering Research*, ISSN 0973-4562, Volume 12, Number 17.

[16]. Ajad Patel, Sonali Gandhi, SwethaShetty,Prof. BhanuTekwani Jan -2017, "Heart Disease Prediction Using Data Mining", *IRJET*, Vol. 4, Issue1.

[17]. https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/

[18]. (https://towardsdatascience.com/logistic regression, ANN

[19]. https://en.wikipedia.org/wiki/Exploratory_data_analysis

[20]. DrSaravanakumar, Eswari, Sampath, Lavanya "Predictive Methodology for Diabetic Data Analysis in Big Data," *Elsevier, ISBCC*, 2015.

[21]. Stephanie Revels, Sathish A.P. Kumar and Ofir Ben-Assuli, "Predicting Obesity Rate and Obesity-Related Healthcare Costs using Data Analytics", *Health Policy & Tech.*, http://dx.doi.org/10.1016/j.hlpt, 2017.

[22]. Vijayalakshmi N, UmaMaheswari M., "Data mining to elicit predominant factors causing infertility in women", *IJCSMC*, Vol. 5, Issue. 8, August2016.

[23]. Min Chen, YixueHao, Kai Hwang, Lu Wang and LigWang, "Disease prediction by machine learning over big data from Healthcare communities", *IEEE Access*,2017.

[24]. Yu, C.Y.; Li, X.X.; Yang, H.; Li, Y.H.; Xue, W.W.; Chen, Y.Z.; Tao, L.; Zhu, F., "Assessing the Performances of Protein Function Prediction Algorithms from the Perspectives of Identification Accuracy and False Discovery Rate",*Int. J. Mol. Sci*., 2018.

[25]. Menzies, N.A.; Wolf, E.; Connors, D., "Progression from latent infection to active disease in dynamic tuberculosis transmission models: A systematic review of the validity of modelling assumptions", *LancetInfect. Dis*., 2018.