Ensemble Learning: Predicting Human Pathogenicity of Hematophagous
Arthropod Vector-Borne Viruses

Huakai Hu¹,², Chaoying Zhao³,¹, #, Meiling Jin¹, Jiali Chen⁵, Xiong Liu¹, Hua Shi¹, Jinpeng Guo¹, Changjun Wang³,¹, *, Yong Chen¹, ², * ⁵

¹School of Public Health, China Medical University, Shenyang, Liaoning province, China
²China Chinese PLA Center for Disease Control and Prevention, Beijing, China
³School of Public Health, Zhengzhou University, Zhengzhou, Henan Province, China
⁴Liaoning Provincial Center for Disease Control and Prevention, Shenyang, Liaoning, People’s Republic of China
⁵School of Medicine, NanKai University, Tianjin, People’s Republic of China.

Running title: Predicting Pathogenicity of Hematophagous Arthropod Vector-Borne Viruses

*These authors contributed equally to this work and are listed as co-first authors

Correspondence to Yong Chen chenyonger@126.com
Changjun Wang, science2008@hotmail.com

Abstract

Hematophagous arthropods occupy a pivotal role in ecosystems, serving as vectors for a wide array of pathogens with significant implications for public health. Their capacity to harbor and transmit viruses through biting actions creates a substantial risk of zoonotic spillover. Despite the advancements in metagenomic approaches for virus discovery in vectors, the isolation and cultivation of viruses still pose significant challenges, thereby limiting comprehensive assessments of their pathogenicity. Here, we curated two datasets: one with 294 viruses, characterized by 37 epidemiological features, encompassing virus information and host associations; the second with 71,622 sequences of hematophagous arthropod vector-borne viruses, annotated with 33 sequence features. Two XGBoost models were

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
developed to predict arbovirus human pathogenicity—one integrating macroscopic eco-epidemiological data, the other incorporating virus-related sequence features. The macroscopic model identified non-vector host transmission as a key determinant, especially involving Perissodactyla, Artiodactyla, and Carnivora Order. The sequence-based model demonstrated that viral adhesion and viral invasion had distinct trends with consistent increase and decrease in the likelihood of virus pathogenicity to humans, respectively. With validated through an independent dataset, the model exhibited a congruous alignment with documented pathogenicity outcomes. Together, the models offer a holistic framework for assessing the pathogenic potential of viruses transmitted by hematophagous arthropods.

Introduction

Hematophagous arthropods, such as mosquitoes and ticks, play a pivotal role in ecosystems as blood consumers and crucial disease vectors (Cuthbert et al., 2023; Touray et al., 2023). These arthropods can harbor a myriad of pathogens, including bacteria, fungi, and viruses. Notably, viral infections in these organisms are classified under the umbrella terms of Arthropod-Borne Viruses (arboviruses) and insect-specific viruses (ISVs) (Calisher & Higgs, 2018; Gould et al., 2017; Nouri et al., 2018; Zhao et al., 2022). The potential for these arthropods to harbor and disseminate a diverse array of pathogens poses a grave threat to both human and animal health, with the ominous potential to trigger outbreaks and result in a substantial number of annual fatalities (Batson et al., 2021; Roth et al., 2018). Vector-borne diseases contribute significantly to infectious diseases (Chala & Hamde, 2021), with notable arboviruses including the Zika virus (Khongwichit et al., 2023; Weaver et al., 2018), Japanese encephalitis virus (JEV) (Kampen & Werner, 2014), and the incessant menace of Dengue virus (DENV) (Fournet et al., 2023).

In recent years, propelled by the widespread adoption of Viral Metagenomics sequencing technologies, the identification of a wide range of established and emerging viruses within hematophagous vectors, such as mosquitoes and ticks, has become feasible (Ni et al., 2023; X. Yang et al., 2023). This technological progress presents an unprecedented opportunity to
comprehensively explore the distribution and transmission patterns of arboviruses and ISVs across a spectrum of hosts, including both vectors and non-vectors. Such advancements are crucial for supporting early warning systems, facilitating the anticipation and mitigation of disease spread before its onset (Birnberg et al., 2020; Brinkmann et al., 2016). Despite these achievements in viral metagenomics, current bioinformatic methods for virus recognition still face limitations (Fang et al., 2019). Accurate identification of a significant number of unknown contigs remains challenging. Even when identifying known or novel viruses, the direct isolation and cultivation of these viruses from vectors proves to be formidable tasks, hindering in-depth exploration of their pathogenesis and immune response (Lewis et al., 2021).

In general, the close phylogenetic relatedness among viruses can offer insights into their potential for human infectivity, as closely related viruses are generally presumed to share common phenotypes and host ranges (Geoghegan & Holmes, 2018). However, despite being a common rule of thumb for virus risk assessment, the extent to which evolutionary proximity to viruses with known human infectivity accurately predicts zoonotic potential remains unexamined in the current literature (Behl et al., 2022). Furthermore, the specific model is designed to be trained on sequence features of closely related viruses (i.e., strains of the same species) to discern viruses with human infectivity (Zhang et al., 2019). Unfortunately, this method often overlooks critical functional characteristics of the viral genome, resulting in a model that is less inclined to identify universally applicable pathogenic features across diverse viruses. Consequently, predictions derived from such a model may be highly susceptible to substantial biases (Mollentze et al., 2021).

The epidemiological characteristics of virus transmission encompass not only the virus itself and information about its vector host (Zaid et al., 2021; Y.-J. S. Huang et al., 2019a; Viglietta et al., 2021) but also factors such as geographical and climatic variations, as well as interactions with non-vector hosts (Ciota & Keyel, 2019; Conway et al., 2014; Forrester et al., 2014; Tabachnick, 2016). Moreover, for specific viruses, their nucleotide sequence information may reflect actual pathogenic details (Bartoszewicz, Genske, et al., 2021). Therefore, through a comprehensive analysis that integrates both macroscopic and microscopic perspectives, our objective is to identify the epidemiological features and viral
sequence characteristics that have the greatest impact on the potential pathogenicity to humans.

Based on the global data of arthropod-borne virus compiled by Huang et al. (Y. Huang et al., 2023) as a foundation, we carefully curated the contents to extract pertinent information concerning hematophagous arthropod-borne viruses. Additionally, to augment our analysis, we utilized SeqScreen for insightful functional details of the viral sequences (Balaji et al., 2022). Employing the XGBoost algorithm with ensemble learning, we developed both regression and classification prediction models. This facilitated the identification of factors with the most significant impact on human pathogenicity and enabled the construction of ensemble learning for predicting the pathogenicity of virus sequences carried by hematophagous arthropods.

Materials and methods

Database restructuring and epidemiological feature retrieval

The initial dataset comprised 101,094 virus sequences sourced from NCBI, spanning the period from March 11, 1991, to January 28, 2023 (Y. Huang et al., 2023). To enhance the reliability and specificity of our analysis, a stringent screening process was applied, systematically excluding records lacking host information, sampling location details, and those with ambiguous vector-host relationships. It is important to note that this dataset excludes data from Antarctica. Subsequently, we identified 11 species of hematophagous arthropods, including mosquitoes and ticks (Table supplement 1), while excluding non-blood-feeding species such as Tipulidae and Chironomidae. Following this, we systematically screened the entire dataset, retaining records exclusively related to hosts classified as hematophagous arthropods. This refined dataset, derived through meticulous curation, forms the foundation for our research, ensuring the integrity and accuracy of subsequent analyses. Nevertheless, due to inaccuracies in the classification of vectors within the database, a Python script was developed to scrape taxonomic directory. This script retrieved detailed order, family, and genus information for each hematophagous arthropod and non-vector host classification. To reveal the distinct composition of non-vector hosts, host
counts underwent logarithmic transformation (Figure 1A). For a more comprehensive presentation, host classifications with fewer than 100 occurrences were amalgamated into an “others” category, resulting in a total of 10 host classifications (Figure 1C).

In terms of additional epidemiological features, Köppen climate classification data for each vector were acquired based on their discovery locations. This information was sourced from both the Weather and Climate website (https://weatherandclimate.com/) and the Mindat website (https://www.mindat.org/). Concurrently, continental data for each country were obtained from the World Population Review (https://worldpopulationreview.com/continents) and Baltimore classification data were sourced from the International Committee on Taxonomy of Viruses (ICTV) (https://ictv.global/report/genome).

The development of a regression model for macroscopic characteristics

Firstly, among the 8,468 datasets in this study, we employed an R script to transform it into a dataset comprising 294 distinct virus types, each characterized by 37 unique features (Table 1). Subsequently, we utilized the XGBoost ensemble learning model to establish regression models. The dataset was divided into training and validation sets at a ratio of 7.5:2.5. Given the balanced ratio of positive to negative samples (1:1) in the model’s database, addressing imbalance was not deemed necessary. The training set was employed to train the model based on the specified parameters (Table supplement 2). After determining the optimal number of iterations through 10-fold cross-validation, we proceeded to construct the final model using this identified count.

Development and validation of a macroscopic features classification model

In the microscopic pathogenicity classification model, we annotated the aforementioned database uniformly using SeqScreen, resulting in a total of 71,622 virus sequences. After excluding viruses from hosts submitted to NCBI after 2022, we obtained a final dataset of 71,593 sequences for this model. Due to the imbalance in positive and negative samples in the database (positivity rate of 79.3%), we adjusted the sample sampling rate to balance the dataset. The specific parameters employed in this model are meticulously detailed in Table supplement 3. Utilizing the training set, the model was trained in accordance with these parameters, determining the optimal iteration count through rigorous 5-fold cross-validation. Subsequently, the final model was constructed utilizing the identified optimal iteration count.
To constitute an additional validation dataset, we retrieved Ebinur Lake Virus with arthropods as hosts from NCBI, incorporating these samples with those previously excluded. Following a consistent application of the specified parameters, we trained the model using functional features from the entire dataset. Subsequently, predictions were generated on the additional validation dataset. The obtained results underwent a meticulous comparative analysis with findings from established pathogenic studies.

Results

Global overview of hematophagous vector-virus distribution, diversity, and host interactions

This study has curated a comprehensive dataset of 8,468 hematophagous vector-virus pairs, shedding light on their geographical distribution, diversity, and interactions with hosts. In terms of distribution, these vectors were classified into two principal classes: Insecta and Arachnida, spanning seven distinct families (Figure 1A). The records cover all six continents except Antarctica, spanning across 102 countries globally and representing 24 diverse climate types (Figure 1B). Regarding diversity, among the hematophagous vectors, Culicidae (64%, 5,445 sequences) predominates, constituting over half of all records, followed by Ixodidae (32%, 2,703 sequences). Globally, the United States exhibits the highest diversity and abundance of vectors, hosting five distinct families, followed by China with four. In terms of virus records associated with vectors, the United States (1,977) leads the list, followed by Russia, China, and Japan.

Turning to non-vector hosts, the dataset includes an additional 54,789 pairs of non-vector hosts and viruses, with the non-vector hosts categorized into 15 groups. Among these hosts, humans are the most prevalent, accounting for 40,078 records, followed by Artiodactyla and Aves, constituting nearly 20% of the total (Figure 1C). The interactions between viruses and non-vector hosts are distinct. The majority of viruses are associated with a single host. Notably, West Nile virus (WNV) and Tick-borne encephalitis virus (TBEV), both belonging to the Flaviviridae, exhibit the most widespread cross-host transmission, being detected in
nine non-vector host species. Moreover, as viruses expand their capacity to infect a wide range of non-vector hosts, a noticeable reduction in viral diversity is observed. Specifically, viruses capable of infecting only one host encompass 10 distinct virus families, while those exhibiting infectivity across two to four hosts are confined to five families. Remarkably, viruses with the ability to infect five, six, or seven hosts are prominently represented by families such as Flaviviridae and Togaviridae. Among viruses capable of infecting seven hosts, Dabie bandavirus stands out as a unique case. Belonging to the Phenuiviridae, this virus is predominantly found in Asia (China, Japan, and South Korea). Infection with Dabie bandavirus poses a significant health risk, causing a severe febrile illness accompanied by thrombocytopenia, known as Severe Fever with Thrombocytopenia Syndrome (SFTS), leading to its alternate nomenclature as the SFTS virus. The Japanese encephalitis virus exhibits the highest degree of cross-vector host diversity, being detectable in three distinct vector families: Culicidae, Ixodidae, and Ceratopogonidae. The previously mentioned WNV and TBEV demonstrate transmission capabilities across both vector families, Culicidae and Ixodidae.

Pathogenicity of hematophagous arthropod vector-borne viruses: a macroscopic regression analysis of epidemiological characteristics

Through transforming the mentioned database and enhancing it with additional epidemiological characteristics, we constructed a comprehensive dataset for the model. This dataset comprises 294 distinct viruses, each characterized by 37 diverse features, broadly categorized into viral characteristics, vector host features, and non-vector host features. To unpack the crucial factors underlying human pathogenicity, we constructed and rigorously trained an XGBoost model. This model leverages human infection status as the dependent variable and incorporates 36 diverse features as independent variables, pinpointing the key determinants of human infection. The model exhibits robust performance on the testing set, with minimal prediction errors reflected in low MSE (0.01) and MAE (0.05) values, highlighting its accuracy. Additionally, high R² (94.20%) and Explained Variance (94.29%) values underscore the model’s comprehensive ability to explain the variance in the dependent
variable.

The detection of viruses in non-vector hosts significantly influences human pathogenicity (Figure 2), surpassing the impact of both vector hosts and the viral agents themselves. Notably, the characteristics “Cross_host”, representing the total diversity of non-vector hosts in which the virus has been detected, carries a weight of 52 in the model. This underscores the critical role of the diversity of non-vector hosts in determining human pathogenicity. Specifically, when considering potential human-pathogenicity, the order of importance is as follows: Perissodactyla, Artiodactyla Carnivora and Aves. The higher the diversity of virus detections in these animals, the greater the likelihood of the virus being pathogenic to humans.

After non-vector host factors, the subsequent important set of characteristics relates to the vector hosts. Among these, “Cross_vector_g”, representing interspecies transmission among diverse vector genus, emerges as the most critical factor. If a virus can propagate within diverse vector genus, there is a substantial likelihood of viral spillover. The third set of characteristics relates to the intrinsic characteristics of the virus itself. The closer the viral phylogenetic relationship, the higher the likelihood of inducing similar immune responses, thereby leading to diseases.

**Relationship between viral genomic function and human pathogenicity: a microscopic machine learning approach**

In our research, we employed SeqScreen to functionally annotate all viral sequences in our comprehensive dataset. After excluding sequences without successful annotations, our refined dataset comprised 71,622 arboviruses and ISV sequences, each accompanied by their respective functional, host, and pathogenic features. The largest category within our dataset consists of mosquito-borne arboviruses, with Dengue virus 1 (9,194 sequences), Dengue virus 2 (8,999 sequences), and West Nile virus (4,656 sequences) being the most prevalent. Tick-borne arboviruses, including African swine fever virus (3,915 sequences) and Crimean-Congo hemorrhagic fever orthonairovirus (3,771 sequences) closely follow in quantity.

Our functional annotation revealed a total of 10 distinct pathogenic features. The results indicated that “viral adhesion” is the most prevalent function, accounting for 62% (44,482
sequences). This function facilitates virus adhesion to host cells, initiating infection and paving the way for subsequent invasion and replication. Following closely are the “viral counter-signaling” (49%) and “host xenophagy” (47%), which are typically associated with immune evasion. These mechanisms enable the virus to survive, replicate within host cells, and successfully transmit to other cells (Table 1).

Among known non-pathogenic viruses to humans, “viral invasion” stands out as the most prevalent function, despite its relatively lower overall count compared to other functions. Notably, within these viruses, the primary hosts targeted are hematophagous arthropod vectors, with non-vector hosts predominantly represented by Artiodactyla and Aves (Figure 3A). Conversely, within the known human-pathogenic viruses, “viral adhesion” ranks as the most prevalent function in terms of annotation quantities. In this context, excluding human hosts, hematophagous arthropod vectors continue to be predominant, followed by Aves (Figure 3B). This observation suggests a potential genomic similarity in the pathogenicity of arboviruses among humans, hematophagous arthropod vectors and Aves.

We developed a binary classification XGBoost model using 33 features, which included functional annotations for all viruses in the database and viral size (length of virus). The model’s dependent variable denotes whether a virus is pathogenic to humans. After excluding the viruses in the extra validation dataset, the remaining viruses in the database were allocated to a training set and a testing set in a 7.5:2.5 ratio. While achieving a high accuracy (95.36%), we incorporated additional metrics, such as Precision (97.57%), Recall (96.55%), and F1 score (97.06%), for a more nuanced assessment. Furthermore, we generated an ROC curve (Figure 4A) and a confusion matrix (Figure 4B) to gain a holistic view of the model’s strengths and weaknesses.

The model’s results clearly demonstrate that, in terms of average gain, “viral adhesion” exhibits the highest value, significantly enhancing the model’s predictive accuracy. “Host xenophagy” and “viral invasion” closely follow suit (Figure 5A). Regarding the model’s coverage, “viral invasion” and “host ubiquitin” occupy the top two positions due to their capability to impact a wide array of viral sequences (Figure 5B). In terms of the model’s weights, the size of the viral sequence takes precedence over other features, indicating its frequent utilization in the model construction and its vital role in making supplementary
assessments on the virus pathogenicity based on functional insights (Figure 5C).

In our conclusive analysis, we employed SHAP (SHapley Additive exPlanations) to gain deeper insights into the individual feature contributions to the model. Notably, the top-ranking feature—“viral size”—does not exhibit a discernible trend in pathogenicity to humans. However, other features reveal intriguing patterns. Specifically, both “viral adhesion” and “host xenophagy”, although slightly less significant than size, individually demonstrate distinct trends: viral sequences annotated with either of these functions consistently increase the likelihood of virus pathogenicity in humans. Conversely, “viral invasion” demonstrates an inverse relationship, wherein sequences possessing this trait tend to reduce the probability of virus pathogenicity. The majority of remaining features, on the other hand, positively correlate with pathogenicity. In summary, most features contribute towards determining the likelihood of virus pathogenicity in humans (Figure 6).

To delve deeper into the intricate interactions among these features, we conducted a thorough analysis. Our results highlight that among all features, “viral counter signaling” exhibits the most significant interaction with viral size. However, its impact on pathogenicity remains inconclusive, lacking a definitive directional trend (Figure 7A). Additionally, we observed a noteworthy interaction between “host xenophagy” and “viral adhesion”. The concurrent presence of these features substantially enhances the virus's pathogenicity towards humans (Figure 7B). Interestingly, “viral invasion” demonstrates a strong but contrasting interaction with “viral counter signaling”. Specifically, “viral counter signaling” seems to function as a protective factor against human pathogenicity when “viral invasion” is present, leading to a reduced likelihood of the virus being pathogenic to humans (Figure 7C).

In our comprehensive analysis of interactions across all features, significant insights emerged. While viral size lacks a clear discernible trend in its interaction with other features, the interplay of 'host xenophagy' with both “viral adhesion” and 'viral counter signaling' guides the model toward non-pathogenicity predictions, acting as a protective feature (Figure S1).

To assess real-world performance, we compiled an additional dataset consisting of 29 viruses carried by hematophagous vectors, submitted after 2022. This dataset comprises 24 strains of SFTS virus, 3 of Restan viruses, 3 of Tataguine viruses, 1 of Japanese encephalitis virus (JEV), and 1 of Nairobi sheep disease virus (NSDV). Furthermore, 25 sequences of Ebinur
Lake virus (EBIV) borne by arthropods were downloaded from the NCBI, resulting in a total of 54 virus sequences. To ensure independence, the content of the additional dataset was excluded from the original model database. The train and test datasets were merged to train the ultimate model. Validation was then conducted using the additional dataset. Model predictions indicated that all sequences of SFTSV, a specific strain of JEV, a particular variant of Tataguine virus, and one isolate of Ebinur Lake Virus potentially exhibit pathogenicity to humans.

Discussion

The intricate interplay between hematophagous arthropods and the viruses they harbor forms a dynamic ecosystem with profound implications for public health. Ticks and mosquitoes, integral components of ecosystems, serve as potent disease vectors capable of transmitting various pathogens, including arboviruses and ISVs. Recent advancements in Viral Metagenomics sequencing technologies have significantly transformed our investigation of the virome within hematophagous vectors, providing unparalleled access to comprehensive viral genetic information. Although technological advancements have been made, bioinformatic methods for virus recognition still face inherent limitations, particularly in identifying unknown contigs, which hinders comprehensive virome characterization. The isolation and cultivation of known or novel viruses from vectors remains challenging, impeding in-depth exploration of their pathogenesis and immune response. The common approach to assessing pathogenicity relies on viral phylogenetic analysis, assuming that viruses with significant phylogenetic distance share similar pathogenic properties. However, the extent to which phylogenetic relatedness accurately predicts the potential for zoonotic diseases remains a critical aspect requiring deeper exploration. Existing models, often tailored for closely related viruses, may inadvertently oversimplify crucial functional characteristics, introducing biases in predictions. To address these challenges, this study innovatively adopts an ensemble learning algorithm in machine learning. Our objective is to comprehensively explore the human-pathogenicity of viruses borne by hematophagous vectors, considering
both macro and micro-level characteristics, with the ultimate aim of identifying key viral
features associated with pathogenicity.

Based on the curated dataset, the distribution and abundance of hematophagous arthropods
suggest that the United States, Russia, and China harbor the highest number of vector insects,
notably mosquitoes and ticks, acting as primary carriers for arthropod-borne viruses (Y.-J. S.
Huang et al., 2019b; Wu et al., 2023). Among the viruses carried by these arthropods, RNA
viruses, specifically those classified as dsRNA under the third Baltimore group, dominate.

*Phlebotominae* and *Ceratopogonidae* share virus families, while distinct mosquito genera
harbor a diverse array of viruses, primarily belonging to the *Flaviviridae*, *Togaviridae*, and
*Peribunyaviridae* (Figure supplement 1). The abundance of viruses is influenced by various
factors, prompting a correlation analysis on the dataset. The results reveal strong correlations
for most viruses, excluding *Asfarviridae*, Zirqa virus, and Wallerfield virus, with six key
characteristics, including vector family and weather conditions. Within the community of
vector-borne viruses, *Flaviviridae*, *Togaviridae*, *Bunyaviridae*, *Rhabdoviridae*, and
*Phenuiviridae* are commonly co-detected and considered as core virome (Coatsworth et al.,
2022) (Figure supplement 2).

Within the cyclic dynamics of arboviruses, numerous factors intricately influence the
transmission to humans, impacting pathogenic outcomes. Variations in the intrinsic nature of
viruses yield diverse levels of human pathogenicity, commonly associated with phylogenetic
proximity. Moreover, an increased diversity of viruses within vectors may foster co-infection,
thereby facilitating viral evolution and spill-over (Vogels et al., 2019). Consistent with these
observations, our results highlight the significant influence of viral intrinsic factors on human
pathogenicity. These viruses primarily propagate diseases through hematophagous vectors,
predominantly mosquitoes, the species and behaviors of these vectors, collectively termed
“Vector capacity” (Conway et al., 2014), along with environmental shifts (Hermanns et al.,
2023; Weissenböck et al., 2010), play a pivotal role in shaping vector composition and
consequentially impact viral transmission. Notably, our research reveals that the impact of
vector hosts is equivalent to that of viral intrinsic factors. Both the diversity across vector
genera and the quantity of vector genera exert a substantial influence on human pathogenicity.
Interactions between viruses and non-vector hosts drive viral evolution (Sen et al., 2016),
with interspecies interactions being the primary driver for viral spill-over (Y.-J. S. Huang et al., 2019a). The interplay between vectors and Aves hosts enables long-distance viral transmission (Forrester et al., 2014), while interactions with vertebrates also emerge as pivotal determinants (García-Romero et al., 2023; Golnar et al., 2014; Stephenson et al., 2019). Our results align with these insights, highlighting that, beyond interspecies categories, viruses infecting Perissodactyla and Artiodactyla pose the most significant risk for human pathogenicity, increasing the likelihood of transmission to humans and subsequent disease outbreaks.

Capitalizing on the ability to analyze viral genomic data for predicting viral pathogenicity to humans (Bartoszewicz, Seidel, et al., 2021), we conducted an exhaustive investigation into the genetic functionalities of arboviruses and ISVs within the database. Utilizing the ensemble learning algorithm, we meticulously developed and trained a predictive model. Furthermore, an additional dataset comprising arboviruses submitted to NCBI after 2022 was incorporated for validation and prediction purposes. Our model demonstrates superior performance, showcasing distinctive contributions from individual functional features that collectively shape the overarching trend in viral pathogenicity. Specifically, “viral adhesion”, representing a pivotal mechanism for viral infection and entry into host cells, emerges not only as the predominant feature but also significantly enhances the overall performance of the model. Empirical evidence affirms that the presence of this feature in viral sequences, subsequent to transmission to humans by hematophagous vectors, consistently indicates an elevated risk of pathogenicity. For instance, viruses within the Flaviviridae, such as DENV, WNV, and ZIKV (Begum et al., 2019; Cruz-Oliveira et al., 2015; Faustino et al., 2019; Hasan et al., 2017; Martins et al., 2019), utilize E and capsid proteins to enter receptor cells. Likewise, the Chikungunya virus, a member of the Togaviridae, facilitates the fusion with receptor cells through trimeric E1/E2 spikes (Ciota & Keyel, 2019).

While “viral invasion” plays a pivotal role in the initial phase of viral entry, it is notably scarce among this dataset (Table 1). Furthermore, most viruses with this feature are presently classified as non-pathogenic to humans (Figure 3). The high abundance of Flavivirus in the dataset could explain the limited occurrence of “viral invasion”, given their unique infection mechanisms that may not require this specific feature. Additionally, SeqScreen might not
have detected “viral invasion” in these sequences. Despite its relatively low occurrence in the
dataset, this feature exhibits the highest cover value in the XGBoost model, indicating its
significant impact on the model’s performance. Interestingly, contrary to expectations, its
presence predominantly acts as a “protective factor” in predicting pathogenicity, as revealed
by SHAP explanations. This phenomenon may be attributed to the prevalence of this feature
among non-pathogenic viruses in our dataset. However, since we balanced the samples during
model training, it could also be a result of interactions among different features. A more
in-depth exploration of interactions related to this trait revealed a robust interplay with “viral
counter signaling”. When both features coexist, the model significantly leans towards
predicting non-pathogenicity in humans. Importantly, these two processes are not mutually
exclusive factors in actual virus infections. This observation implies potential distinctive
invasion mechanisms of arboviruses, indicating unconventional pathways for entering host
cells that facilitate immune evasion.

In this dataset, “viral counter signaling” and “host xenophagy” are prevalent features actively
enhancing virus pathogenicity and triggering host infection. They play a crucial role in the
pathogenicity to humans (Costa et al., 2013; King et al., 2020). Notably, “host xenophagy”,
similar to “viral adhesion”, significantly influences the model results. In terms of interactions,
it has the strongest interaction with “viral adhesion”, leading to a positive inclination towards
pathogenicity.

The feature of “size”, representing the length of viral sequences, while not directly associated
with pathogenic functions, plays a crucial role in refining the final results, as indicated by the
model’s weight (Figure 5C). Training the model with only 33 functional features resulted in
unreliable accuracy (82%) and a high false-positive rate. However, the inclusion of “size”
substantially improved the model’s performance. Notably, the influence of “size” on viral
pathogenicity lacks a discernible trend (Figure 6), resulting in predictive outcomes that tend
towards a more stochastic distribution. Interaction analysis revealed that “viral counter
signaling” has the strongest interaction with “size” (Figure 7). Even with these interactions,
determining the direction of pathogenicity remains challenging. In summary, “size’ appears to
fine-tune the model’s final predictions. When combined with functional features, it facilitates
a more accurate assessment of the likelihood of pathogenicity to humans.
In the validation results using an additional dataset, we identified four viruses with the potential to infect humans and induce diseases. Firstly, all sequence of Dabie bandavirus were predicted to be pathogenic. These viruses collected from ticks in Miyazaki Prefecture, Japan, exhibited a high degree of homology through phylogenetic analysis with a virus previously isolated from an SFTS patient, providing strong evidence for its potential pathogenicity (Sato et al., 2021). The model also predicted the pathogenicity of a strain of JEV, detected in mosquitoes in the Qinghai-Tibet region of China (Li et al., 2011). Despite the elevated altitude of the region, the presence of antibodies against the virus in both the indigenous population and swine suggests a localized occurrence of virus transmission, thereby challenging the initial presumption that the virus would not be prevalent at higher elevations.

Tataguine virus (Kapuscinski et al., 2021), isolated from Anopheles sp. in Gambia, belongs to the Peribunyaviridae. While its pathogenicity to humans remains inconclusive, there is a high likelihood of infection symptoms if transmitted through hematophagous vectors. Among the 25 strains of Ebinur Lake Virus, one isolated from Hyalomma marginatum in the Volgograd region of Russia in 2023 was predicted to be capable of infecting humans. This virus, commonly found in China's prevalent vector host, Culex modestus, has been studied extensively for its ability to infect BALB/c mice, resulting in pronounced clinical symptoms (Zhao et al., 2020). Recent studies have substantiated the capacity of Aedes aegypti to serve as a vector for such viruses (C. Yang et al., 2022). Although antibodies have been detected in human serum samples, the lack of positive RT-PCR results prevents a conclusive determination of the virus's ability to infect humans and induce diseases (Xia et al., 2020). This aligns with the model's prediction, as the six viruses isolated from Culex modestus, included in the model, are unlikely to be pathogenic to humans.

This study endeavors to leverage machine learning methodologies for discerning overarching factors influencing the pathogenicity of hematophagous vector-borne viruses in humans. Our developed predictive model, focused on gene function, has successfully demonstrated the capability to predict virus pathogenicity in humans. However, it is crucial to acknowledge certain limitations in our study. In the global dataset of vector-borne viruses, there exists an uneven distribution, particularly with an overabundance of viruses such as DEV. This imbalance may result in an unavoidable bias that impacts the accuracy of the model.
Furthermore, the selected machine learning algorithms, while effective, may not match the
efficacy of neural networks, posing challenges in optimizing for the current abundance of data.

Notably, variations in blood-feeding habits among hematophagous vectors were not
considered, which can significantly contribute to the spread of viruses. Different vector
species may exhibit distinct preferences and behaviors in their blood-feeding patterns,
influencing the transmission dynamics of viruses. Future research should incorporate these
behavioral nuances to provide a more comprehensive understanding of virus dissemination. In
summary, our model provides a novel perspective and serves as a valuable tool for the further
analysis of virus sequences, providing effective information for the monitoring and early
warning of hematophagous arthropod vector-borne transmission.

In this investigation, our primary objective is to discern both macroscopic and microscopic
factors influencing the risk of human pathogenicity in hematophagous vector-borne viruses.

Employing ensemble learning standpoint, we uncovered key characteristics associated with
viral pathogenicity from an epidemic perspective. Simultaneously, we delved into pivotal
functional features impacting human pathogenicity at a molecular level, with a specific focus
on the functional aspects of viral sequences. Moreover, we deploy our developed model to
forecast the human infectivity of viral sequences within an additional validation dataset. The
model's performance in predicting the pathogenicity of these viruses at the genetic level not
only enriches our comprehension of established and emerging virus risks but also broadens
the scope of hematophagous arthropods detection. Importantly, it contributes substantively to
the mitigation of present and future risks associated with vector-borne diseases.

Acknowledgement

The authors acknowledge the global open dataset shared by Huang et al and Xuan Li for
assistance with additional data collection. The laboratory is funded by a grant from National
Key Research and Development Program of China (2019YFC1200501). The funders had no
role in study design, data collection and interpretation, or the decision to submit the work for
publication.
Data availability

The data supporting the findings of this study are available upon reasonable request from the author. Researchers interested in accessing the dataset for further exploration or verification are encouraged to contact Huakai Hu at hhyu98@163.com for assistance. We are committed to promoting transparency and collaboration in scientific research, and we welcome inquiries regarding the data underlying our published results.

Author contribution

Huakai Hu, Idea Generation, Data Curation and Transformation, Model Development and validation, Visualization, Writing - original; Chaoying Zhao, Conceptualization, Methodology, Writing - original draft and review; Jiali Chen, Conceptualization, Methodology, Writing - review and editing; Meiling Jin, Conceptualization, Writing – review and editing; Hua Shi, Conceptualization, Writing – review and editing; Jinpeng Guo, Project administration, Writing – review and editing; Changjun Wang, Conceptualization, Methodology, Writing – review and editing; Yong Chen, Supervision, Funding acquisition, Project administration, Writing – review and editing;
References


Cruz-Oliveira, C., Freire, J. M., Conceição, T. M., Higa, L. M., Castanho, M. A. R.


Sato, Y., Mekata, H., Sudaryatma, P. E., Kirino, Y., Yamamoto, S., Ando, S., Sugimoto...
o, T., & Okabayashi, T. (2021). Isolation of Severe Fever with Thrombocytopenia Syndrome Virus from Various Tick Species in Area with Human Severe Fever with


CC-BY-NC 4.0 International license

It is made available under a CC-BY-NC 4.0 International license.
Table 1: Summary of Epidemiological characteristics in regression model. A detailed summary of the 37 epidemiological characteristics considered in our regression model.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Name of characteristics</th>
<th>Detailed description of characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>virus</td>
<td>Virus_Group</td>
<td>Arboviruses or ISVs</td>
</tr>
<tr>
<td>virus</td>
<td>virus name</td>
<td>virus genus</td>
</tr>
<tr>
<td>vi_G</td>
<td>virus genus</td>
<td></td>
</tr>
<tr>
<td>vi_F</td>
<td>virus family</td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>virus counts</td>
<td></td>
</tr>
<tr>
<td>baltimore</td>
<td>virus baltimore</td>
<td></td>
</tr>
<tr>
<td>vector hosts</td>
<td>vector_G</td>
<td>vector genus</td>
</tr>
<tr>
<td></td>
<td>vector_F</td>
<td>vector family</td>
</tr>
<tr>
<td></td>
<td>vector_O</td>
<td>vector order</td>
</tr>
<tr>
<td></td>
<td>vector_C</td>
<td>vector class</td>
</tr>
<tr>
<td>continent</td>
<td>vector continent</td>
<td></td>
</tr>
<tr>
<td>country</td>
<td>vector country</td>
<td></td>
</tr>
<tr>
<td>climate</td>
<td>vector climate</td>
<td></td>
</tr>
<tr>
<td>cross_vector_G</td>
<td>Counts of cross-vector</td>
<td>Counts of cross-vector genus</td>
</tr>
<tr>
<td>cross_vector_F</td>
<td>counts of cross-vector</td>
<td>Counts of cross-vector host families</td>
</tr>
<tr>
<td>cross_vector_O</td>
<td>vector orders</td>
<td>Counts of cross-vector host orders</td>
</tr>
<tr>
<td>cross_vector_C</td>
<td>counts of cross-vector</td>
<td>Counts of cross-vector host classes</td>
</tr>
<tr>
<td>total_vector</td>
<td>Total counts of</td>
<td>Total counts of cross-vector hosts</td>
</tr>
<tr>
<td>vector_G_T</td>
<td>cross-vector host genera</td>
<td></td>
</tr>
<tr>
<td>vector_F_T</td>
<td>cross-vector host families</td>
<td></td>
</tr>
<tr>
<td>vector_O_T</td>
<td>cross-vector host orders</td>
<td></td>
</tr>
<tr>
<td>vector_C_T</td>
<td>cross-vector host classes</td>
<td></td>
</tr>
<tr>
<td>non-vector hosts</td>
<td>Aves</td>
<td>Aves host</td>
</tr>
<tr>
<td></td>
<td>Carnivora</td>
<td>Carnivora host</td>
</tr>
<tr>
<td></td>
<td>Rodentia</td>
<td>Rodentia host</td>
</tr>
<tr>
<td></td>
<td>Chiroptera</td>
<td>Chiroptera host</td>
</tr>
<tr>
<td></td>
<td>Primates</td>
<td>Primates host</td>
</tr>
<tr>
<td></td>
<td>homo</td>
<td>homo host</td>
</tr>
<tr>
<td></td>
<td>Didelphimorphia</td>
<td>Didelphimorphia host</td>
</tr>
<tr>
<td></td>
<td>Artiodactyla</td>
<td>Artiodactyla host</td>
</tr>
<tr>
<td></td>
<td>Perissodactyla</td>
<td>Perissodactyla host</td>
</tr>
<tr>
<td></td>
<td>Eulipotyphla</td>
<td>Eulipotyphla host</td>
</tr>
<tr>
<td></td>
<td>Reptilia</td>
<td>Reptilia host</td>
</tr>
<tr>
<td></td>
<td>Lagomorpha</td>
<td>Lagomorpha host</td>
</tr>
<tr>
<td></td>
<td>Anura</td>
<td>Anura host</td>
</tr>
<tr>
<td></td>
<td>Pirola</td>
<td>Pirola host</td>
</tr>
<tr>
<td></td>
<td>Diprotodontia</td>
<td>Diprotodontia host</td>
</tr>
<tr>
<td>cross_host</td>
<td>total cross non-vectors</td>
<td>total cross non-vectors host</td>
</tr>
<tr>
<td>homo_infected</td>
<td></td>
<td>Whether or not homuns are infectious</td>
</tr>
</tbody>
</table>
**Table 2: Summary of FunSoCs annotation results from SeqScreen.** Counts and definitions of 10 distinct FunSoCs identified in this dataset.

<table>
<thead>
<tr>
<th>FunSoC title</th>
<th>Counts</th>
<th>FunSoC definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viral adhesion</td>
<td>44482</td>
<td>Mediates viral adherence to host cells</td>
</tr>
<tr>
<td>Viral counter signaling</td>
<td>35256</td>
<td>Viral suppression of host immune signaling within host cells to avoid inflammatory responses</td>
</tr>
<tr>
<td>Host xenophagy</td>
<td>33656</td>
<td>Target host xenophagy/autophagy</td>
</tr>
<tr>
<td>Viral invasion</td>
<td>4376</td>
<td>Mediates viral invasion into host cell</td>
</tr>
<tr>
<td>Host transcription</td>
<td>949</td>
<td>Target host transcription to inhibit or activate</td>
</tr>
<tr>
<td>Host ubiquitin</td>
<td>880</td>
<td>Target host ubiquitination machinery</td>
</tr>
<tr>
<td>Host cell death</td>
<td>802</td>
<td>Target host apoptotic cell death pathways either to inhibit or activate</td>
</tr>
<tr>
<td>Resist complement</td>
<td>144</td>
<td>Enable resistance from host complement components</td>
</tr>
<tr>
<td>Antibiotic resistance</td>
<td>4</td>
<td>Counters the effect of antibiotics administered to inhibit the growth or vital functioning of bacterial or eukaryotic parasites.</td>
</tr>
<tr>
<td>Induce inflammation</td>
<td>1</td>
<td>Directly activate host inflammatory pathways to cause damage</td>
</tr>
</tbody>
</table>
Figure 1: Hematophagous arthropod vector and non-vector hosts characteristics in the dataset. (A) The global distribution and quantity of blood-sucking vectors and their carriers. (B) The number of vector hosts, the continents where they are located, and the types of hostile weather conditions where they are found. (C) The characteristics of the number of non-vector hosts. (D) Viruses transmitted across non-vector hosts Quantity. The abscissa is the number across non-vector hosts, and the ordinate is the total number of viruses.
Figure 2: Relative importance of different macroscopic characteristics in the regression model. The weight contributions of diverse epidemiological features of viruses in the regression model to human pathogenicity.
Figure 3: Hosts distribution in viral functions annotation. The distribution of hosts for known non-pathogenic viruses to humans (A) and known human-pathogenic viruses (B). The actual counts of viruses are converted to percentage representations in their respective sections of the chart.
Figure 4: Metrics of comprehensive assessment of model performance. The utilization of ROC Curve (A) and Confusion Matrix (B) for assessing the performance of the model.
Figure 5: Ranking of metrics presented in the XGBoost functional annotations model.

Within the results of the XGBoost model, the functional feature importance outcomes of gain (A), cover (B), and weight (C) are separately obtained. These three metrics collectively reflect the relative significance in determining the pathogenicity of hematophagous arthropod vector-borne viruses.
Figure 6: The collective impact of viral function annotations on pathogenicity prediction

analyzed through SHAP. Providing a comprehensive overview of how various viral function
annotations collectively contribute to the model's predictions regarding pathogenicity.
Figure 7: Detailed analysis of the interactions among crucial features in pathogenicity prediction models through SHAP. The interactions examined include those between viral sequence size and viral adhesion (A), host xenophagy and viral adhesion (B), as well as viral invasion and viral counter signaling (C). These analyses contribute to a deeper understanding of the combined influence of these features on pathogenicity predictions.
Supplementary Information

Table supplement 1: Family for hematophagous arthropod vector

<table>
<thead>
<tr>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culicidae</td>
</tr>
<tr>
<td>Phlebotominae</td>
</tr>
<tr>
<td>Ceratopogonidae</td>
</tr>
<tr>
<td>Simuliidae</td>
</tr>
<tr>
<td>Tabanidae</td>
</tr>
<tr>
<td>Cimicidae</td>
</tr>
<tr>
<td>Ixodidae</td>
</tr>
<tr>
<td>Argasidae</td>
</tr>
<tr>
<td>Stenoponiidae</td>
</tr>
<tr>
<td>Phthiraptera</td>
</tr>
</tbody>
</table>

Table supplement 2: Hyperparameter settings for the XGBoost regression model.

Optimized parameter settings for the XGBoost regression model obtained through rigorous experimentation and fine-tuning.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>booster</td>
<td>dart</td>
</tr>
<tr>
<td>eta</td>
<td>0.15</td>
</tr>
<tr>
<td>max_depth</td>
<td>3</td>
</tr>
<tr>
<td>subsample</td>
<td>0.7</td>
</tr>
<tr>
<td>objective</td>
<td>reg:logistic</td>
</tr>
<tr>
<td>tree_method</td>
<td>exact</td>
</tr>
<tr>
<td>max_cat_threshold</td>
<td>20</td>
</tr>
<tr>
<td>eval_metric</td>
<td>&quot;logloss&quot;, &quot;auc&quot;, 'error&quot;</td>
</tr>
</tbody>
</table>
Table supplement 3: Hyperparameter settings for the XGBoost classification model.

Optimized parameter settings for the XGBoost classification model obtained through rigorous experimentation and fine-tuning.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>objective</td>
<td>binary:logistic</td>
</tr>
<tr>
<td>tree_method</td>
<td>exact</td>
</tr>
<tr>
<td>scale_pos_weight</td>
<td>0.26</td>
</tr>
<tr>
<td>eta</td>
<td>0.15</td>
</tr>
</tbody>
</table>