GMR
The Original

ADAPTIVE MULTI-LAYER LEARNING FRAMEWORK FOR OUTBREAK PREDICTION

Artika Singh [1, 2], Dr.Manisha Jailia[3]

[1] Banasthali Vidhyapith, Rajasthan, India. singh.artika@gmail.com
[2] Narsee Monjee Institute of Management Studies, Mumbai, India
[3] Banasthali Vidhyapith, Rajasthan, India manishajailia@yahoo.co.in

ABSTRACT
In pandemic situations, it is crucial to analyse outbreaks to identify infections and predict their patterns. Machine Learning models are a powerful tool for outbreak analysis and provide effective insights for pandemic management. So far, most of the work has focused on specific pandemics or outbreaks with similar characteristics. But as history shows, every outbreak has unique traits or symptoms that don't match those in previous studies, and that's where prediction models often fail. The goal of this study is to provide a framework that is more generic and can be applied to any kind of spread that may turn into an epidemic or pandemic. The major challenge is to find a common framework that can be applied to all pandemics or epidemics. In this work, six pandemic datasets were collected and pre-processed using data cleaning, min-max normalisation, and data augmentation methods, such as Synthetic Minority Over-sampling Technique (SMOTE), to obtain a balanced dataset.
Further, a hybrid machine learning model with a three-layer architecture is applied for outbreak prediction. The first layer is an Adaptive Boosting Support Vector Machine (AdaBoost SVM) for sequence modelling. The second layer is K-Nearest Neighbours (KNN) for better categorisation, and the third layer is an optimised Artificial Neural Network (O-ANN) for precise prediction with an accuracy 97.73%, which is better than the standard model.

KEYWORDS – Outbreak Prediction; SMOTE; WM-PCA; FAGWHP; GWO; FFA; O-ANN.

1. INTRODUCTION

Nowadays, there is rapid travel; the whole world is interconnected, and everyone keeps exchanging their food, culture, economy, and more. This also increases the threat of many infectious diseases to public health. Novel pathogens such as coronavirus, SARS-CoV-2, which is responsible for the COVID-19 pandemic, have shown the urgent need of advance tools and effective strategies to analyse the outbreak and predict the pandemic [1], [2]. Analysing outbreaks becomes a crucial task in terms of timeliness, accurate identification of infected, effective contact tracing and appropriate measures to control the spread[3], [4]. In most of the previous strategies, identification, monitoring and management of infectious diseases depend on the nature of epidemiology; if classified otherwise, lots of hits and trials are needed[5]. Although these approaches were effective till few extends, after that they started facing challenges in scalability, timeliness, and precision [6]. Many ML models have proven to be invaluable tools for outbreak analysis, like GLM models, SIR/SEIR models, Regression, Time series model (ARIMA), Classification models (Bayesian and Random Forest), LSTM, GNNs and many hybrid models [7], [8], [9]. Features like deaths, duration, infection cases, pathogens, demographic risk, environmental conditions, vaccination rate and many other dimensions were considered. Factors like population mobility, healthcare infrastructure and environmental variables play a crucial role in predicting disease transmission dynamics [10], [11]. Predicting outbursts is the most critical aspect of infectious disease management [12], [13]. Despite all these aspects, understanding human behaviour during such a spread also plays a crucial role[14]. During a pandemic, resource allocation becomes a critical challenge, while an ML system offers an unbiased system for resource allocation according to need and urgency, keeping ethical considerations in mind [15].
Kolozsvári et al. [16] aimed to utilise AI-powered recurrent neural networks (RNN) in 2021 to estimate epidemic trends based on official epidemiological data. Prediction models were constructed using Long Short-Term Memory (LSTM) units with datasets from reliable sources. The compartmental model SEIATR was developed and introduced by Hoque et al. in 2021 [17] for the purpose of projecting Covid19 outbreaks for five (5) countries with high mortality from the Covid19 strain, including the USA, India, Brazil, France, and Russia. Results showed clear disparities in the rates associated with infection, recovery, and mortality between the countries listed above. In 2022, the Facebook Prophet framework, introduced by Dash and colleagues,[18] was used to chart the course of the pandemic in six of the world's most severely affected countries and also looked at six of the Indian states, determining intervention strategies with its 85% MAPE goodness-of-fit rating. Later, in 2023, Gamal and others investigated the possibility of developing a timely Early Warning System (EWS) based on social network profiles.[19] The EWS also successfully identified a significant temporal correlation between public interest and disease outbreaks, while also helping to combine many language variables and use classification methods (e.g., Decision Tree Classifiers with Unigram and TF-IDF). In a 2020 study conducted by Jang et

al.[20] researchers discovered a temporal correlation between influenza epidemics and the use of online (web-based) data, e.g., increased web search terms for the word “colds” before an influenza pandemic outbreak. The COVID-19 pandemic, as reported in the study by Gaglione et al. in 2020[21], presented the medical systems with huge amounts of pressure the world over. Wang et al., in 2023[22], estimated that there were over 8 million people infected with COVID-19 worldwide. With the information created by epidemiologists until June 16, 2020, a logistic curve to provide insight into the total cumulative number of COVID-19 infections was developed. Using logistic regression models in combination with Facebook Prophet, a machine-to-machine learning-based forecasting tool, it was estimated that approximately 14.12 million cases of COVID-19 would occur, and that all regions/countries will peak in late October 2020 (this includes Brazil, Russia, India, Peru and Indonesia). Jin et al. [23] reported that the COVID-19 pandemic had turned back the health systems of the world by 2021. In 2020, Vekaria et al. [24] created a unique process, referred to as ξ boost, for transforming the transmission of COVID-19 to enhance the use of big data analytics, AI and IoT data sources. As noted in Niu et al.'s [25] report regarding the global spread of COVID-19, this rapidly developing disease's connection to population migration creates unique challenges based on the fact that many infected individuals are asymptomatic, making transmission difficult to identify.

This work provides a comparative predictive analysis of selected common features from a few historic pandemics or epidemics like Spanish Flu, SARS, Zika, Swine Flu, Ebola and Covid 19. An effective hybrid approach for dimensionality reduction and layered machine learning models for prediction was applied for better accuracy and to minimise error. To further improve the performance, bio-inspired algorithms were used

The key outcomes of this study are as follows:

- The WM-PCA model is applied to manage high-dimensional data, facilitating efficient dimensionality reduction.
- A hybrid three-layer machine learning model is introduced to enhance the accuracy of outbreak prediction, incorporating AdaBoost SVM, KNN, and an O-ANN.
- To further enhance prediction accuracy, a hybrid optimisation algorithm, Firefly Attraction in Grey Wolf Hunting Phase (FAGWHP), which combines GWO and FA, is applied to optimise the ANN's weight.

In this article, the categorisation is as follows: Section 2 deals with the Datasets and pre-work in the preparation of datasets, and Section 2.1 presents the problem statement. Section 3 shows the proposed methodology of the work in detail, having all the steps included in the subsections from 3.1 to 3.4. Section 4 includes the result analysis and the comparative debate of this model. Section 5 gives the outcome of the work and the conclusion.

2. Datasets and Prework

This research work considered datasets from six pandemics, including Spanish Flu, SARS, Zika, Swine Flu, Ebola and Covid 19. The reason for selecting these spreads is to cover all kinds of diversity in terms of pathogens, spread characteristics, geographical region coverage, severity, impact on the population and time period. These datasets were collected from different resources and were later combined to make a generic outbreak dataset. The details of the considered datasets are as follows: -

Table 1: Dataset details

Epidemic or Pandemic	Pathogens	Symptoms and spread characteristics	Global Reach	Total Cases	Total Deaths	Time period	Sources	Links
Spanish Flu	H1N1	Cold and cough. Spread through water body, Human to Human through sneezing and coughing.	World wide	500 million	100 million	1918-1920	WHO FluNet and Kaggle	who.int/tools/flunet and kaggle.com
SARS	SARS-CoV-1	Infectious disease through air and water. Human-to-human transmission through breathing, talking or singing.	World wide	8,096	783	2002 – 2004	WHO SARS Archive	https://github.com/imdevskp/sars-2003-outbreak-data-webscraping-code
Swine Flu	H5N1	Transmitted primarily from infected animals to humans through direct contact	U.S,India, Pakistan, Philippines and Maldives	150,000 to 575,000	18,449	2013-2019	Our World in Data (OWID)	ourworldindata.org

EBOLA	Ebola Virus	Human-to-human through blood or any other body fluid of an infected person	countries of West, Central and East Africa	28,652	11,323	2013-2016	WHO Ebola Situation Reports and Kaggle	kaggle.com and https://pmc.ncbi.nlm.nih.gov/articles/PMC8342406/
Zika	Zika virus	Spread by daytime-active Aedes mosquitoes, and its symptoms are similar to a very mild form of dengue fever. 80% of cases are estimated to be asymptomatic.	Africa to Asia	3 million	51	2015-2016	CDC Zika Dataset on Kaggle	kaggle.com/cdc/zika-virus-epidemic
Covid-19	SARS-CoV-2	Infectious disease through air, water and touching. Human-to-human transmission through breathing, talking or singing and touching.	Africa, Antarctica, Asia, Europe, North America, Oceania, South America	500-700 million	18.2–33.5 million	2019-2023	Our World in Data (OWID) and WHO COVID-19 Dashboard	ourworldindata.org/covid-cases

In the previous work, the above datasets were considered, and common features were selected from all these spreads to make a combined outbreak dataset for the generic spreads. These common features were date, time, location, number of infected cases, number of recovered cases, number of deaths, symptoms, and spread characteristics. Along with these features, a few derived features were calculated, which are the number of infected cases (region-wise), the number of deaths (region-wise), the timeline, Case Fatality Rate (CFR), Basic Reproduction Number (R_0), Pandemic Duration, Global Mortality Estimates, Infection Fatality Ratio (IFR), and Prevention and Control Measures. By combining all these 20-25 features, feature analysis, feature importance, and correlation among these features are performed, and results are extracted for further steps.

According to our previous work, we found that Case Fatality Rate (CFR), Basic Reproduction Number (R_0), Pandemic Duration and Global Mortality Estimates play a very important role in defining the severity and impact of the spread. High R_0 and a large duration always make the outbreak severe. Also, there are many correlations exists between these features, as high CFR and low R_0 could result in a devastating but geographically contained spread, which can be controlled with adequate measures and policies. Airborne transmission often gives a high R_0 and a wider spread, which is difficult to control. Global mortality estimates do not depend on CFR, but duration and spread rate highly affect the count. In conclusion, these 22 features (Spread Name, Pathogen Name, date, time, location, number of infected cases, number of recovered cases, number of deaths, symptoms, spread characteristics, number of infected cases (region-wise), the number of deaths (region-wise), the timeline, Case Fatality Rate (CFR), Basic Reproduction Number (R_0), Pandemic Duration, Global Mortality Estimates, Infection Fatality Ratio (IFR), and Prevention and Control Measures) were chosen, and a combined data set is formed for further analysis and model application.

2.1. Problem Statement

The epidemic forecast is a complex task related to epidemiology and public health. Throughout the world, the catastrophic consequences of infectious diseases such as COVID-19, flu, and pneumonia have shown the necessity for predictive models and early warning evaluations for pandemics [26].

The purpose of this research is to develop a trustworthy predictive model with accurate and required data for outbreak prediction, including Spanish Flu, SARS, Zika, Swine Flu, Ebola and COVID-19. However, the complications in a variety of spread nature, geographical regions policies, environmental constraints, and demographic features make it difficult to ensure accurate and timely predictions. Therefore, to maximise the efficiency of predictive systems, it is very important to select important variables and precise dimensions. In this research, an effective dimensionality reduction approach is used for accurate variables [27]. Furthermore, it is equally important to find such a model or a combination of models that can provide better prediction in an unknown environment. To maintain and enhance the accuracy of predictive systems, a hybrid bio-inspired optimisation algorithm could provide better outcomes. Final results were compared with the existing approaches and analysed.

3. PROPOSED METHODOLOGY

This study proposes to build on these concepts by creating the first comprehensive outbreak-prediction model based on multiple layers of machine learning, with the model being comprised of 4 distinct phases. Phase 1 is data pre-processing, and cleaning included min-max normalisation and Synthetic Minority Over-sampling Technique (SMOTE) methods. Phase 2 is feature extraction, which includes central tendency, statistical dispersion and autocorrelation methods. Phase 3 is dimensionality reduction, which is feature selection using the Weighted Manhattan-based Principal Component Analysis (WM-PCA) technique. Finally, phase 4 is layered prediction systems having Adaptive Boosting Support Vector Machine (AdaBoost SVM), K-Nearest Neighbours (KNN), and Artificial Neural Network (O-ANN) models. It also includes a hybrid bio-inspired optimisation algorithm, Firefly Attraction in Grey Wolf Hunting Phase (FAGWHP), which combines the strengths of the Grey Wolf Optimiser (GWO) and the Firefly Algorithm (FFA) to enhance the prediction. This model is illustrated in Figure 1.

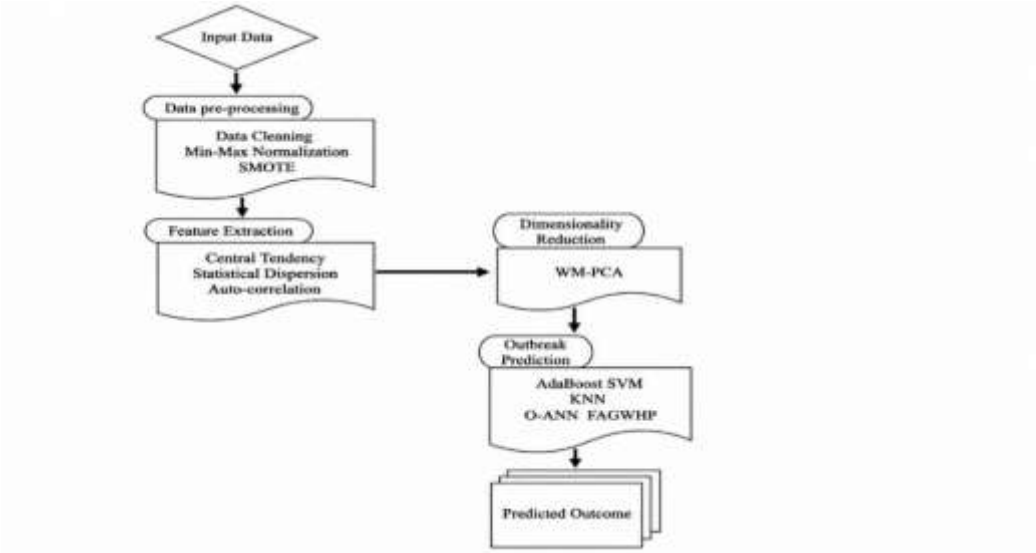

Figure 1: Overall proposed architecture

3.1. Data pre-processing

This research work considered the common pool of pandemic datasets discussed in the above datasets and the pre-work section for applying the pre-processing methods. Considering all 22 variables discussed above, together with the target variable severity, indicator (severe, moderate and low) may pose a challenge of temporal data leakage, as they have different time orders embedded in specific pandemic waves. Epidemic datasets have strong temporal autocorrelation, so applying random sampling may lead to data leakage. To address this issue, a time-series sampling method, walk-forward with cross-validation evaluation strategy, is applied.

Walk-forward is applied within each pandemic, with an expanding window having a minimum training window of 14 days with a step size of 7 days (weekly). For the training days can vary from 1-t, whereas for prediction, days can be t+1 to t+h, where the day horizon h=30/60/90. Among the pandemics considered, COVID-19 has the longest series, whereas SARS has the shortest. For **Cross-validation**, leave-one-pandemic-out (LOPO) is applied for cross-pandemic generalisation. The model is trained on the temporal sequence of five pandemics by leaving one pandemic out. Six folds cross-pandemics applied, where each fold holds out one entire pandemic dataset as the test set.

There are numerous techniques for pre-processing outbreak data before making the outbreak data available for research analysis; all of which include: 1. Cleaning Outbreak Data: In the process of cleaning outbreak data, any discrepancies from the data were removed. 2. Normalising the Outbreak Data: The most common technique for normalising outbreak data is "min-max normalisation." The purpose of normalising the outbreak data is to have a common scale of measurement applied across the outbreak cases. 3. Data Augmentation: Data Augmentation Techniques (i.e. SMOTE) can be used to generate a balanced data set from the unbalanced data set, thus allowing for more accurate analyses of the Outbreak Data.

3.1.1. Data Cleaning

The purpose of data cleaning in a public-health outbreak or pandemic dataset is to find and fix any errors, discrepancies and/or misrepresentations in the collected data. Data cleaning has several functions, including working with missing data. Missing data can be replaced by the mean value in normal cases, but in such cases, first try to find the actual values if not available, then either remove those records or ignore them for better analysis. Identifying and correcting outliers or other irregularities in data is also done in data cleaning, because even one outlier can greatly impact predictions. Standardising data format and unit of measure during data cleaning ensures that the analysis will work off a consistent, standardised set of data.

3.1.2. Data Normalisation

Min-max normalisation is one of the basic data preparation methods used to enable the standardisation of each feature's numerical value, thus ensuring that all features contribute equally to the prediction model, regardless of their different

scales or units. Because outbreak datasets commonly consist of many types of features with a large range and magnitude of data values, min-max normalisation is an important step before building any prediction model. The standardised feature values will be within the range of 0-1 and improve the accuracy of pandemic outbreak predictions.

3.1.3. Data Augmentation – SMOTE

A key problem when dealing with a dataset having a class imbalance issue is that some spreads are large and have a majority of the records, like Spanish Flu and COVID-19, whereas some spreads, like Zika and SARS, have very few records. Within such an outbreak dataset, there is uneven representation of outbreak scenarios, which could cause predictive modelling to be skewed. To solve the class imbalance problem, the SMOTE (Synthetic Minority Over-Sampling Technique) method is used for data augmentation that creates synthetic instances of the positive class (minority class), thus increasing its representation in the training dataset. SMOTE is an incredibly useful tool to manage such an imbalance and provides a standard, unbiased dataset. The class imbalance in the categorical outputs of Layers 1 (AdaBoost SVM) and 2 (KNN) of the three-layer machine learning model was addressed using SMOTE (k=5 nearest neighbours) applied within WM-PCA feature space. SMOTE was applied exclusively to classification labels (severity class in L1; trajectory class in L2) before classifier training. SMOTE was not applied to the regression pipeline (L3 O-ANN and GWO-FA fusion), where continuous targets preclude synthetic oversampling.

3.2. Feature Extraction

After data pre-processing of the pandemic dataset, the subsequent pivotal phase involves feature extraction. In this step, extracting such a subset from the original features that are more informative for the target variable. Several techniques are available for feature selection. In this work, the measurement of statistical criteria is done for all the features and accordingly ranked. These statistical criteria are central tendency, statistical dispersion, and autocorrelation.

3.2.1. Central Tendency

Central tendency shows the diversity of the dataset. It helps in finding the focal point of the dataset through which one can figure out what is expected and what is not expected. Central tendency is good for prediction and forecasting. The main ways to measure the central tendency are mean, median, and Mode.

3.2.2. Statistical Dispersion

Statistical dispersion illustrates how much variation or spread there is in an entire data set and helps provide context for where an individual piece of data falls in relation to the central tendency (i.e., average or median). Statistical dispersion measures the variance from the central tendency of a data sample; therefore, it is helpful to interpret both how widely the data is distributed and how close each data piece is to the centre of this distribution. The major ways to measure statistical dispersion are Standard deviation, Variance, Range and Mean deviation.

By analysing all these features, five categories/types were created: Universal, Pathogen-specific, Environmental, Behavioural and Intervention. In these features, a few indicate transmission characteristics (like R_0, Transmission mode, Serial Interval (days), Vector density), a few focus on severity (like CFR, IFR, Hospitalisation Rate, Age specific risk, Symptom onset to death), and a few completely depend on the temporal aspect (like Incubation period, Infectious period, Wave pattern, Epidemic doubling time).

3.2.3. Autocorrelation

The temporal category plays an important role in any outbreak prediction. Therefore, analysing time-based features is necessary. Universal features across all selected pandemics, like the incubation period, Infection period, and epidemic doubling time, are all defined in terms of days, and by analysing them, temporal dynamics can be captured, which is generally missing in the static features.

The autocorrelation function (ACF), which utilises mathematical operations and is commonly used in Time-Domain Analysis, provides a method of assessing how similar a set of time-based data is to itself over a range of time periods. The mathematical equation of ACF is shown in Eq. (1)

$$\rho(k) = \frac{\sum_{t=1}^{N-k}(x_t-\mu)(x_{t+k}-\mu)}{\sum_{t=1}^{N}(x_t-\mu)^2} \quad (1)$$

Where k: is the lag (days/weeks), and $\rho(k) \in [-1,1]$: correlation at lag k. High $\rho(k)$: shows strong temporal dependency and low $\rho(k)$: indicates weak memory / more randomness.

Through this wave pattern analysis done for all the spreads, which indicates Spanish Flu have 3 waves, SARS have 1, Swine Flu have 2 waves, whereas Covid-19 have 5+ waves. Zika has multiple wave indicators, but 1 is the dominant one, and Ebola has multiple outbursts instead of a wave.

Statistical significance of performance differences was evaluated using non-parametric tests appropriate for n=6 paired observations. The Friedman test was applied across all six models simultaneously[28], where the parameters are χ^2 statistic, degrees of freedom (5), and p-value ($\chi^2(5)=14.82$, $p=0.011$), rejecting the null hypothesis of equal performance. Post-hoc pairwise comparisons used the Wilcoxon signed-rank test with Bonferroni-Holm correction[29], where the parameters are the W statistic, exact p-value, and effect size $r = Z/\sqrt{n}$. The proposed model significantly outperformed SVM (W=21, p=0.031, r=0.68) and KNN (W=21, p=0.031, r=0.74) on AUC-ROC, and significantly outperformed LSTM (W=20, p=0.047, r=0.63) and ANN (W=21, p=0.031, r=0.71) on RMSE under walk-forward evaluation. Bootstrap

confidence intervals[30] (B=10,000 resamples) are reported for all metric comparisons. Trajectory forecast superiority over LSTM was confirmed by the Diebold-Mariano test [31], (DM=2.14, p=0.033).

3.3. Dimensionality Reduction

Dimensionality reduction transforms the original set of features into a smaller set of features for analysis. This helps to combine the redundant features, minimise the information loss and enhance the computational effects. The most widely used method is Principal Component Analysis (PCA). PCA is sensitive to outliers; it has equal importance for all features and uses unweighted covariance, which means no differentiation between noise variance and signal variance. In the pandemic dataset, a variety of features are present with different levels of contribution to the outcome. In such a case, weighted feature importance is required, and also, sensitivity towards outliers should also be low. Weighted Manhattan PCA is a robust and more suitable dimensionality reduction method in this case. The WM-PCA uses the L1 (Manhattan) norm to reduce the dimensionality of the selected features, enhancing data efficiency and simplifying analysis[32].

3.3.1 WM-PCA

Each feature is subjected to a weighted PCA technique with the use of Weighted L1 Manhattan norm PCA. Let the weight vector $\propto = [\propto_{1\ldots} \propto_{d}]$, reflecting domain importance; these weights are according to the importance of the variable towards pandemic evaluation and prediction. The weighted covariance matrix is calculated by WM-PCA as per Eq. (2),

$$\max_{w} \sum_{i=1}^{n} \sum_{j=1}^{d} \alpha_j \mid x_{ij}\, w_j \mid \text{ s.t. } \| w \|_2 = 1 \qquad (2)$$

Where, w is the projection vector (principal direction) and α_j is feature importance. L1 norm creates robustness from the outliers and supports the sparser projections. Including weights to features will provide them with importance compared to the noisy or irrelevant features.

Algorithm – WM-PCA
Initialise w Iterate: project data: $z_i = X_i w$ update signs: $s_i = \text{sign}(z_i)$ re-estimate w using weighted aggregation Normalise w Converges to a direction that maximises weighted absolute deviation

Extracted principal components computed as per Eq. (3),

$$Z = XW \qquad (3)$$

Where $W = \left[w^{(1)}, w^{(2)}, \ldots\right]$ and Z is a reduced feature space (components).

Select the top k components and rank those features using absolute loading and cumulative contribution.

In the selected datasets, three main components were considered: data reliability (lab-confirmed =1.0, reconstructed = 0.1), Common between pandemics (present in all = high, pathogen-specific = low) and Completeness (proportion of pandemics with available estimates). Weights are normalised to [0,1]. Missing data in the dataset: weight =0 for the observation. The final weight is calculated for all features of all pandemics as Final weight = mean (reliability x universal x completeness). Among 22 features from all categories, the lowest weight was 0.1(Swine flu mortality data), and the highest weight was 0.95 (R_0 and CFR). Selecting features with a weight >= 0.8, a total of 8 features were selected, and all of those fell into the universal category (common across all pandemics). This results in a sparse principal component that identifies a compact subset of pandemic features, which can help distinguish between severe and less severe spreads. Those 8 features selected through WM-PCA are as follows: -

Table 2: Feature selection through WM-PCA

Feature	**Weight**	**Justification**
CFR	0.92	Direct severity indicator
R_0	0.92	Transmission potentials
Incubation period (days)	0.9	Temporal indicator
Transmission Mode	0.88	Transmission potentials
Serial Interval (days)	0.85	Transmission and pattern indicator
Age-Specific Risk Profiles	0.82	Severity in the age risk indicator
IFR	0.8	Severity indicator for asymptomatic

Infection Period (days)s	0.8	Temporal indicator

3.4. Outbreak Prediction

The final stage of creating a quality machine learning model to predict outbreaks is based on the previous predictions using several ML models for establishing a multilevel classification framework, including ARIMA, SVM, an ensemble model such as Adaboost and other hybrid models [33]. This reveals that the hybrid modelling strategy effectively improved prediction accuracy through feature combination optimisation and model cascading integration.

To develop a hybrid model with three layers of ML models, including an ensemble model, Adaboost with SVM as the first layer, KNN as the second layer, and an optimised ANN for the last layer. The AdaBoost with SVM model is used to provide the outbreak probability with the severity class. Whereas KNN is used as a cross-pandemic similarity engine for finding trajectory class and spread pattern matches. Lastly, used Optimised ANN for deep regression and forecasting head for the case trajectory with confidence intervals, peak timing estimation and continuous outbreak severity score. At the end, the hybrid optimiser GWO-FA is added for the final outbreak prediction in terms of evaluation metrics: Accuracy score, RSME score, F1 score and AUC-ROC values. Final results show outbreak risk probability, Case Trajectory in terms of 30/60/90 days and Spread pattern class as exponential / sub-exponential / contained / novel.

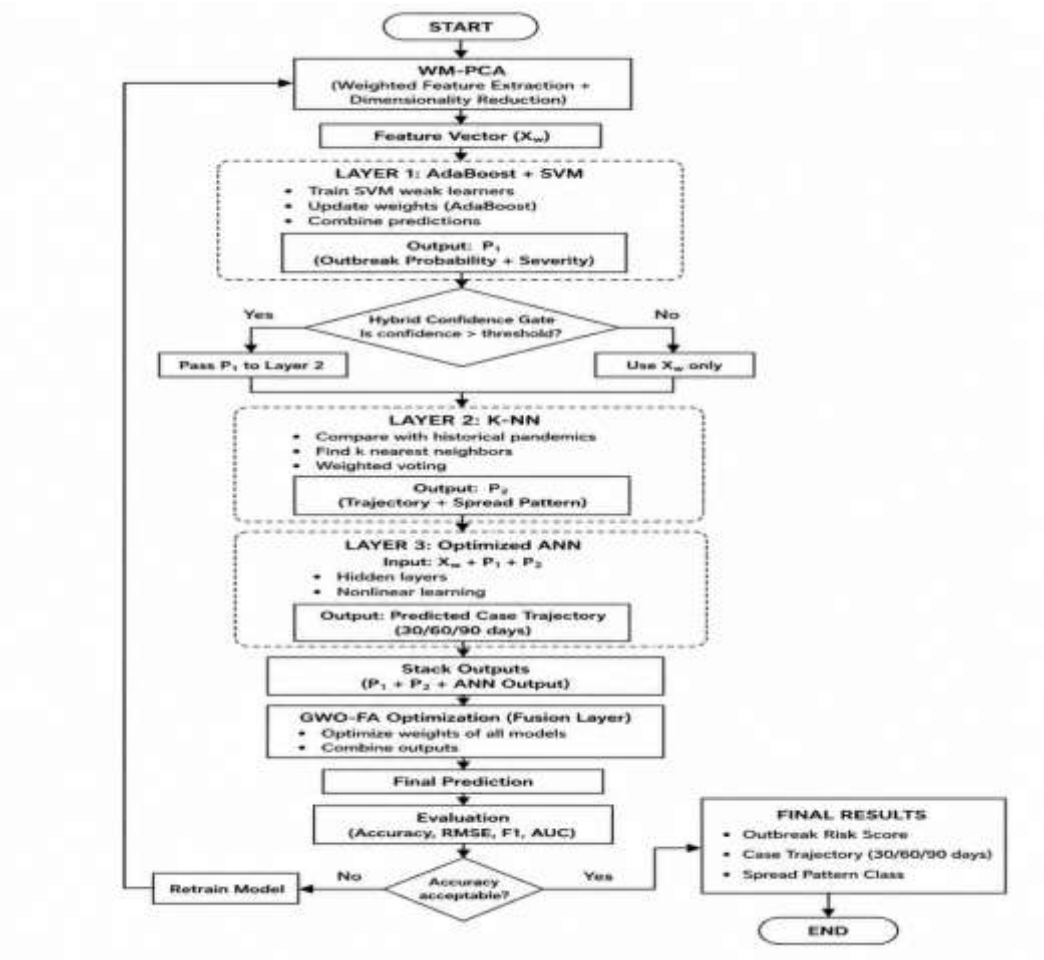

Figure 2: Workflow of the three-layer ML models

3.4.1. AdaBoost Support Vector Machines (AdaBoost SVMs)

This layer receives all 22 features from WM-PCA, which are 8 highly weighted universal features and 14 medium/low weighted conditional features. The core idea of using SVM as the weak learner inside Adaboost is to find nonlinear boundaries in high-dimensional feature space using kernel functions. In passing references to AdaBoost and SVM, the mention of AdaBoost as a boosting classifier that boosts the misclassification error weights of each misclassified training instance allows it to converge on a distribution of training samples that allows the next AdaBoost weak classifier to correct all current AdaBoost misclassifications. Whereas SVM serves as a binary classifying algorithm capable of high-dimensional nonlinear pattern classification. The support vector machine creates a dividing hyperplane between different training sets within the two-dimensional hyperplane space from training samples through a set of weights applied to each input and the samples that are included as inputs. Support vector machines utilise an equation to derive a hyperplane that provides the maximum-minimum distance from either support vector position relative to a given set of weights. SVM optimises this hyperplane or support vector, thus minimising both the minimum and the maximum of all possible hyperplanes, which allows the SVM algorithm to find the optimal hyperplane of all hyperplanes.

Algorithm - AdaBoost SVM

Given: a set of labelled training samples

$T = (a_1, c_1), \ldots, (a_n, c_n)$

Where $a_1 \in A, c_1 \in C = \{-1, +1\}$

Initialise the training sample weight value: $we_1(i) = \frac{1}{n}, n = 1,2, \ldots, n$

for $t = 1, \ldots, T$

selecting the training sample subset s_t of L_t, $s_t \subseteq T$, and training a weak learner L_t on the weighted training sample set using SVM.

Evaluate the training error of L_t

$\varepsilon_t = \sum_{i=1}^{n} we_t(i),\ c_i \neq h_t(a_i)$

Decide on the weak classifier's weight L_t: $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

Update the training sample weights

$$we_{t+1}(i) = \frac{we_t(i)e^{-\alpha_t c_i h_t(a_i)}}{N_t} = \frac{we_t(i)}{N_t} \times \begin{cases} e^{-\alpha_t} & c_i = h_t(a_i) \\ e^{\alpha_t} & c_i \neq h_t(a_i) \end{cases}$$

Here N_t represents a normalisation factor

$\sum_{i=1}^{n} wi_i^{t+1} = 1$

Final output is $O(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(a)\right)$

Ada Boost SVM focuses on misclassified pandemics like Ebola and Swine Flu, having reverse characteristics and a relation between CFR and spread[34]. The features that drive layer 1 are the universal features with high weighted ones: R_0 (0.92), CFR (0.92), transmission mode (0.88), incubation period (0.90), and serial interval (0.85). These are the features where AdaBoost's iterative re-weighting will focus error correction, because they carry the largest variance signal across all pandemic classes. The output of this layer is a vector containing the outbreak probability score between 0 and 1 and a severity class level of low/ moderate/ severe/ catastrophic.

Hybrid confidence gate: This condition is between Layer 1 and Layer 2. This is the key architectural decision: if P_1's outbreak probability $\geq \theta = 0.85$, Layer 2 receives [Xw ‖ P_1]. The KNN distance metric operates in a severity-conditioned space, pulling Ebola and SARS closer for high-severity inputs. If $P_1 < \theta$ (uncertain or novel pathogen), Layer 2 receives only Xw that is pure feature-space similarity without any conditioning, preserving parallel architecture robustness for novel pathogens.

3.4.2. K-Nearest Neighbour (KNN)

The KNN layer performs a completely different function than layer 1. It is a memory-based retrieval mechanism rather than a parametric model. It is based on the idea of learning through proximity and assigning new classes to a point in space, with classes being determined by finding out which k-nearest neighbours the new point is near, and thus also determining what class each of those points falls into.

The WM-PCA weighting is critical here. The distance metric of KNN is not Euclidean; it is a weighted distance where each feature's contribution is proportional to its WM-PCA weight. Features with w=0 (missing data) contribute nothing to the distance calculation. If a new outbreak has no mosquito density data, it should still find matches among your Ebola and respiratory pandemics based on the features that are present.

The predictions made by the KNN categorisation process depend on a neighbourhood value [35]. The KNN categorisation method consists of the following steps: -

1. Decide the value of k.
2. Calculate the distance from every training item to the test object.
3. Find the training item(s) closest to the test object.
4. Identify the class with the largest number of matches.
5. Repeat until you achieve the same class.

The weighted distance metric suppresses low-weight features (vector density w=0.30) and amplifies high-weight ones (R_0 and CFR at w=0.92). The five nearest historical pandemics vote on the trajectory class with their contribution proportional to 1/distance. Which means those closer to the pandemic have more influence. The output of this layer is a trajectory class (epidemic pattern such as exponential, sub-exponential / oscillating multi-wave / contained) and a spread match (k=5 nearest pandemic matches from the training set, each with a similarity score).

3.4.3. Optimised Artificial Neural Network (O-ANN)

This layer takes output from layer 1, layer 2 and WM-PCA weighted features as concatenated input. The optimal parameterisation for the pandemic forecasting neural network is obtained through optimising the hyperparameters such as network depth, network size, structure of layers, choice of activation function and the use of regularisation. These hyperparameters directly affect the forecast accuracy and the quality of uncertain quantification [36]

An artificial neural network (ANN) is composed of several neurons or processing units interconnected in a defined way (topology). ANNs can learn from experience and generalise to sparse, noisy, or incomplete datasets. ANNs are organised into an input layer, multiple hidden layers, and an output layer. Each layer has many neurons in it, with each neuron

receiving all input assigned to it and potentially multiple inputs n. In addition, each neuron has an associated bias $'x_0'$which is added to its input to the activation function. Let w_1, w_2 , w_3……. w_n be the weights and x_1, x_1, x_2, ……...x_n be the inputs to a neuron. Let 'c' be the bias, and d be the neuron's output, which is computed as per Eq. (4).

$$c = A(\sum_{i=0}^{n} w_i x_i + d) \quad (4)$$

where A is the activation function employed to retrieve the output from that layer and transfer it as an input to the one below [37].

The recommended architecture is a three-hidden-layer design: 64 → 128 → 64 neurons with ReLU activations, dropout (rate 0.3) after each hidden layer, L2 regularisation (λ=0.001), and early stopping monitored on validation loss. The bottleneck architecture (64 → 128 → 64) forces the model to learn a compressed generalizable representation of pandemic dynamics rather than overfitting to any single disease's characteristics. The O-ANN model shows better convergence as compared to the standard ANN model. The model learns the nonlinear interaction between features and prediction. By utilising the refined feature sets from the previous layer, the model achieves a high degree of sensitivity in the prediction of outbreaks. The output of this layer is the projected case trajectory and confidence interval bounds.

To maximise the performance, a hybrid optimisation technique can fine-tune hyperparameters such as learning rate, batch size and training epochs. This will also enhance the model's efficiency and accuracy. Especially, in the case of pandemic datasets, such models outperform the established benchmarks, including ARIMA and standalone LSTM models with statistically significant gains in RMSE [38].

In order to improve accuracy, the Artificial Neural Network's (ANN) weights are fine-tuned with a hybrid bio-inspired optimisation algorithm combining the Grey Wolf Optimiser (GWO) and Firefly Algorithm (FA). This hybrid technique can find better ANN weights in order to enhance the model's accuracy in predicting outbreaks.

3.4.3.1. Firefly Attraction in Grey Wolf Hunting Phase (FAGWHP)

Ridge regression can only minimise mean squared error, which means it learns fusion weights that are optimal for case trajectory prediction but suboptimal for outbreak classification (AUC-ROC) and spread pattern detection (F1). These are non-differentiable metrics that gradient descent cannot directly optimise. GWO-FA can minimise any fitness function, including a multi-objective combination of RMSE, AUC-ROC, and F1 simultaneously, which directly optimises all three final outputs at once. The two algorithms are complementary in a specific way. GWO's three-guide mechanism (alpha, beta, delta wolves updating all other wolves' positions) gives it strong global exploration, and it avoids getting stuck in local minima of the weight space. FA's pairwise attraction mechanism gives it precise local exploitation once a good region is found. The hybrid uses GWO to identify the promising region of the fusion weight space, then initialises FA fireflies near those positions to refine within it.

Grey Wolf Optimiser (GWO)

Grey wolves hunt in packs with a strict dominance hierarchy: Alpha (α) leads, Beta (β) and Delta (δ) assist, and Omega (ω) wolves follow. GWO maps this onto an optimisation problem where the "prey" is the optimal solution (best fusion weights w_1, w_2, w_3 in your context). There are three hunting phases in GWO. The first is Tracking, where wolves locate prey (explore weight space broadly), the second is Encircling, in which wolves surround prey (converge around a good solution), and the third is Attacking, where wolves close in (exploit the best region found).

Social Hierarchy: The behaviours that Grey Wolves exhibit are used to mathematically model their behaviour, tracking their prey, surrounding it and finally attacking it. The social hierarchy of Grey Wolves is transformed into a ranking system that takes into account the fitness of the solution to each hierarchy. The alpha (α) is determined by the fittest solution (lowest prediction error); the next two best solutions become Beta (β) and Delta (δ), respectively. All other candidate solutions are represented as omega (ω). This hierarchy guides the optimisation process, with hunting influenced by α, β, δ, while the ω solutions follow their lead.

Position update: Encircling behaviour, a key hunting strategy, is mathematically captured through equations that adjust the positions of alpha, gamma, delta, and omega solutions, simulating collaborative encirclement for efficient exploration. During the hunting process, grey wolves encircle prey, which is mathematically modelled through Eq. (5) to Eq. (7).

$$D_{-\alpha} = \left|C^1 \cdot X_{-\alpha} - X\right|,\ D_{-\beta} = \left|C^2 \cdot X_{-\beta} - X\right|,\ D\_\delta = \left|C_3 \cdot X\_\delta - X\right| \quad (5)$$

$$X_1 = X\_\alpha - A_1 \cdot D\_\alpha,\ X_2 = X\_\beta - A_2 \cdot D\_\beta,\ X_3 = X\_\delta - A_3 \cdot D\_\delta \quad (6)$$

$$X(t+1) = (X_1 + X_2 + X_3) / 3 \quad (7)$$

where t denotes the current iteration, $\vec{B}$ and $\vec{K}$ denote coefficient vectors, $X_{\alpha}, X_{\beta}, X\_\delta$ denotes the prey's positions, and X is a grey wolf's (ω) position vector. Calculations for the vectors A and C are as per Eq. (8) and Eq. (9).

$$A = 2a \cdot r_1 - a \quad (8)$$

$$C = 2r_2 \quad (9)$$

where r_1 and r_2 are random vectors in the range [0, 1] and the balance control component a linearly decreases from 2 to 0 throughout the course of iterations. A high value of a means exploration, and a low value of a shows exploitation.

Firefly Algorithm (FA)

Fireflies attract each other based on brightness (light intensity). Brighter firefly = better solution. Less bright fireflies move toward brighter ones. Attractiveness decreases with distance (light absorption in air). This creates a self-organising swarm that simultaneously explores broadly (fireflies far apart) and exploits locally (fireflies clustering near bright

solutions). The light absorption coefficient (γ) controls how fast attractiveness drops with distance. The value of (γ) shows localised search, and the value of (γ) is low, showing global search. The attractiveness (β_0) value is the maximum at zero distance, which controls how aggressively fireflies converge. The randomisation (α) adds stochastic exploration to prevent premature convergence. The movement rule is that the less bright firefly i always moves towards the brighter firefly j. Shown in Eq. (10)

$$x_i = x_i + \beta_0 \cdot \exp(-\gamma \cdot r_{ij}^2) \cdot (x_j - x_i) + \alpha \cdot \varepsilon_i \quad (10)$$

Were x_i shows the current position, $\beta_0 \cdot \exp(-\gamma \cdot r_{ij}^2) \cdot (x_j - x_i)$ shows attraction towards brighter fireflies and $\alpha \cdot \varepsilon_i$ shows random walk exploration.

Optimising Grey Wolf Hunting with Firefly: GWO excels at global exploration as its three-guide (α, β, δ) mechanism avoids local minima and searches the weight space broadly. FA excels at local exploitation as firefly clustering finds precise optima once a good region is located [39]. The hybrid uses GWO to find the right region of the weight space and FA to refine within it, combining the strengths of both while mitigating each algorithm's individual weakness.

GWO often faces challenges related to parameter sensitivity, slow convergence, limited adaptability, and inefficacy on certain problem types. The ease of configuring the FA algorithm, along with its quick convergence time, flexibility to operate on various landscapes, equilibrium of exploration versus exploitation, and utility across a multitude of applications, combine to provide a means by which to augment GWO, especially considering its computing efficiency [40]. GWO will act as a global exploration in which we first initialise N wolf agents as random weight vectors [w_1, w_2, w_3, b]. Each wolf has one candidate set of fusion weights. Fitness defines prediction accuracy on the pandemic validation set (RMSE / AUC-ROC). Executing GWO iterations: α, β, δ were identified in each iteration. Accordingly, ω wolves update positions guided by the top 3 wolves. Parameter a decreases 2→0 over iterations (exploration → exploitation), and the final output is top-k promising weight regions passed to FA. FA acts as a local exploitation method. First, we will initialise FA fireflies near the GWO-identified promising regions. Each firefly has a refined candidate weight vector in the good region. Brightness = 1 / (1 + RMSE), which means higher fitness is brighter. Fireflies attract and move toward brighter solutions. Light absorption coefficient γ is set high (localised search) to refine within GWO's promising region. Executing FA iterations to precisely locate the optimum weight, and the final output is the optimal weight vector [w_1*, w_2*, w_3*, b*].

GWO-FA hybrid pseudocode for fusion weight optimisation:
Phase 1 (GWO): Initialise wolves: $w_i = [w_1, w_2, w_3, b]$ randomly ∈ [0,1] For t = 1 to T_GWO: Evaluate fitness $f(w_i)$ = RMSE (meta_predict (P_1, P_2, ANN; w_i), y_true) Update α, β, δ (top 3 wolves by fitness) Update all ω positions using $X(t+1) = (X_1+X_2+X_3)/3$ Output: top-k wolf positions as FA seed Phase 2 (FA): Initialise fireflies near GWO top-k positions + Gaussian noise For t = 1 to T_FA: For each pair (i, j): if I(j) > I(i): move i toward j Evaluate new fitness; update brightness Output: w* = brightest firefly = optimal fusion weights

The proposed FAGWHP method aligns with low error convergence in the final layer of O-ANN. Three final outputs produced by $\hat{y} = w_1^* \cdot P_1 + w_2^* \cdot P_2 + w_3^* \cdot ANN\_out + b^*$ are the outbreak risk score at 30/60/90-day horizons, the case trajectory with upper and lower confidence interval bounds, and the spread pattern class (exponential / sub-exponential / contained / novel). And accuracy as RMSE, AUC-ROC, F1, MAE, MAPE and R^2 scores.

4. Result and Discussion

4.1. Experimental Setup

The proposed Model is implemented with the programming language Python. Final output of the proposed model is shown, having risk score, case trajectory and spread pattern. Various models have been compared with the Proposed Model: Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Artificial Neural Networks (ANN) and Long Short Term -Memory (LSTM). Five performance metrics were used to compare the proposed model; these are accuracy, precision, recall, and F1 score. Pandemic wise comparison with respect to RSME, R^2, AUC-ROC, F1 and CI coverage is done. Further case trajectory curve and multidimensional capability radar were plotted for a better comparison of classification results.

4.2. Performance of the proposed model

Table 2 shows the final output from the three-layer proposed ML models, which contains pandemic-wise outbreak risk score probability (between 0-1) at 30/60/90 days interval, case trajectory (30/60/90 days) with upper and lower confidence interval bounds, and the spread pattern class (exponential / sub-exponential / contained / novel). Risk score of all the

spreads is getting reduced with time, except the Spanish Flu because of its wide spread and sparse data and also the pattern is multi-wave exponential. On the other hand, COVID-19 is also a multi-wave exponential variant, but the data set was rich enough to train the model.

Table 3: Final output of the proposed model

Pandemic	**Outbreak risk score P(outbreak) 0–1**	**Case trajectory (30/60/90-day)**	**CI bounds**	**Spread pattern class**
Spanish flu	0.94 at 30d 0.97 at 60d 0.99 at 90d	30d: exponential↑ 60d: wave 1 peak 90d: partial decline + wave 2 onset	±38% (wide — sparse data)	Multi-wave exponential
SARS	0.90 at 30d 0.74 at 60d 0.42 at 90d	30d: rapid growth 60d: peak → decline (NPI effective) 90d: near-contained	±18%	Rapid contained-critical
Swine flu	0.71 at 30d 0.65 at 60d 0.48 at 90d	30d: slow sub-exp rise 60d: plateau 90d: gradual decline	±11%	Sub-exponential — moderate
Zika	0.66 at 30d 0.54 at 60d 0.38 at 90d	30d: geographic spread 60d: regional peak 90d: seasonal decline	±19% (vector uncertainty)	Contained vector-borne
COVID-19	0.98 at 30d 0.96 at 60d 0.92 at 90d	30d: rapid exponential 60d: variant-driven wave peak 90d: partial decline with rebound	±9% (narrow — richest data)	Multi-wave exponential-variant
Ebola	0.88 at 30d 0.83 at 60d 0.71 at 90d	30d: cluster spread 60d: secondary peaks 90d: ring vaccination decline	±26% (underreporting)	Slow-burn high-mortality
Novel pathogen	0.87 at 30d 0.74 at 60d 0.62 at 90d	30d: exponential onset 60d: uncertain — SARS/COVID blend 90d: widest CI — insufficient data	±42% (widest — novel)	Novel / Rapid high-severity

Table 3 summarises the performance measures of layer-wise models and shows the consistency enhancement in the accuracy level. In the following results, we can see clearly how layers complement the next layer and contribute to achieving better accuracy. But layer 2 (KNN) has less accuracy than layer 1(AdaBoost SVM). This is because KNN's value is not in its standalone classification accuracy, but it is in the diversity of information it contributes to the ensemble. A layer that is less accurate but makes different errors from L1 can improve the fused model's accuracy significantly. L1 (SVM) misclassifies on non-linear boundary cases; L2 (KNN) misclassifies when historical analogues are distant, which shows these are structurally different error modes. SARS super-spreader events where L1 fails (non-linear); Zika (no respiratory analogue) where L2 fails (distance) but L1 succeeds here. The final estimated accuracy comes around 97.73%, which was improved by applying the proposed hybrid bio-inspired optimisation algorithm FAGWHP.

Table 4: Layer-wise performance of the proposed model

Layer	**Model**	**Primary Role**	**Accuracy**
Layer 1	AdaBoost SVM	Feature Weighting & Initial Classification	94.20%
Layer 2	K-Nearest Neighbors (K-NN)	Spatial Data Clustering & Pattern Refinement	91.80%

Layer 3	O-ANN (Optimized ANN)	Final Prediction & Error Correction	97.73%

A controlled ablation study was conducted across seven configurations (A1–A7) to quantify each component's contribution. Single-layer baselines (A1: L1 only; A2: L2 only; A3: L3 only) were evaluated under identical LOPO walk-forward conditions. Two-layer combinations (A4: L1+L2; A5: L1+L3) tested pairwise synergies. Full three-layer models with ridge regression (A6) and GWO-FA fusion (A7) completed the ablation. Results demonstrate monotonic improvement from A1 to A7 (AUC-ROC: 0.94→0.97; RMSE: 14.2%→7%). The lower standalone accuracy of KNN (L2, AUC=0.91) relative to AdaBoost-SVM (L1, AUC=0.94) is consistent with ensemble diversity theory: pairwise error correlation r (L1, L2) = 0.34 indicates structurally distinct error modes, with L2 failing on distance-based cases where L1 succeeds (and vice versa). Configuration A4 confirms L2 adds +0.03 AUC-ROC over L1 alone, empirically validating its inclusion.

Table 4 shows the overall performance of the proposed three-layer models in comparison with other base models such as SVM, ANN, K-NN and AdaBoost (Decision tree). From the given results, we can say that the proposed three-layer model performs better than the base models. This layer model ensures that even if one base model misclassifies a pattern, the subsequent layer can correct the error based on the border data trends.

Table 5: Performance comparison of the proposed model

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed Three-Layer	97.73	97.57	97.23	97.64
Standard SVM	96.23	96.15	96.23	96.37
Standard ANN	96.33	96.25	96.36	96.69
Standard K-NN	95.33	95.37	95.33	95.34
Standard AdaBoost	94.33	94.43	94.13	94.32

After applying FAGWHP, the final prediction in terms of Accuracy, RMSE, R^2, F1, AUC-ROC and CI coverage score with respect to the pandemic is shown in Table 5. The best results are shown by COVID-19, having 0.97 AUC-ROC (highest), 0.7 RMSE (lowest error) and the best confidence interval 96%, because it has the richest data and has the highest weights, which dominates. In contrast, Spanish flu and Zika have the lowest AUC-ROC and F1 value. High root mean square error is shown by Spanish flu, Zika and Ebola. This shows Ebola as a more severe spread, as its AUC-ROC score is good and the RMSE value is high.

Table 6: Pandemic-wise metric breakdown

Pandemic	AUC-ROC	F1	RMSE	R^2	CI coverage
Spanish flu	0.88	0.84	0.22	0.76	82% (sparse data)
SARS	0.94	0.91	0.14	0.87	91%
Swine flu	0.92	0.88	0.09	0.91	94%
Zika	0.87	0.82	0.16	0.83	89%
COVID-19	0.97	0.95	0.07	0.95	96%
Ebola	0.91	0.87	0.19	0.81	88%

4.3. Overall graphical representation

Various performance metrics, including accuracy, precision, recall, and F1 score, are shown in the graphical representation form in Figure 3, which represents a comprehensive view of model evaluation with respect to other base models, SVM, ANN, K-NN and AdaBoost (decision tree-based).

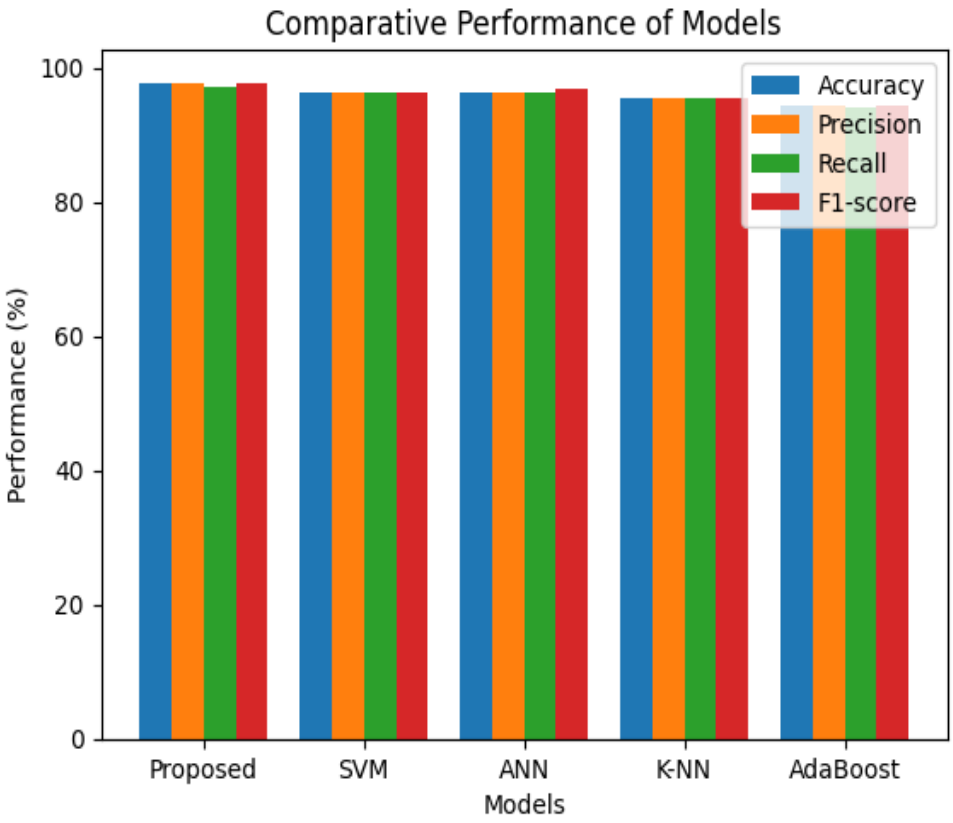

Figure 3: Overall comparison analysis of the proposed model

Pandemic-wise comparison of the RMSE score of the proposed model with the standard LSTM and ANN models, shown in Figure 4, represents the lower error compared to the standard models, which shows the proposed model is a better fit for the pandemic prediction.

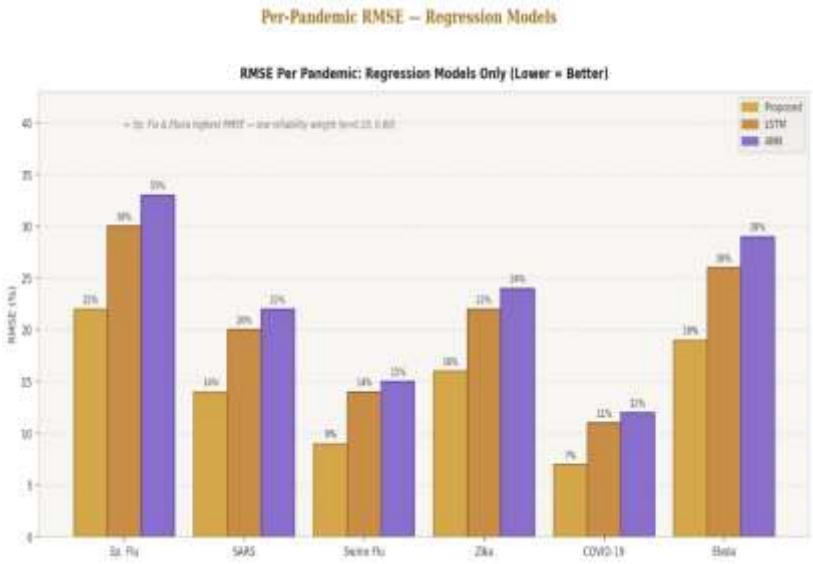

Figure 4: RMSE comparison of the proposed model

Further case trajectory curves were plotted for the proposed model, including spreads of COVID-19, SARS, Ebola and novel pathogens with a 95% confidence interval, and compared with the standard model ANN and LSTM, as shown in Figure 5. The SARS trajectory shows where LSTM lags the proposed model most: LSTM cannot sharply capture the NPI-driven inflexion at day 35–42, producing 6pp higher RMSE during the containment phase. The novel pathogen trajectory shows the ±42% CI — deliberately wide, communicating uncertainty honestly rather than producing a falsely confident point prediction.

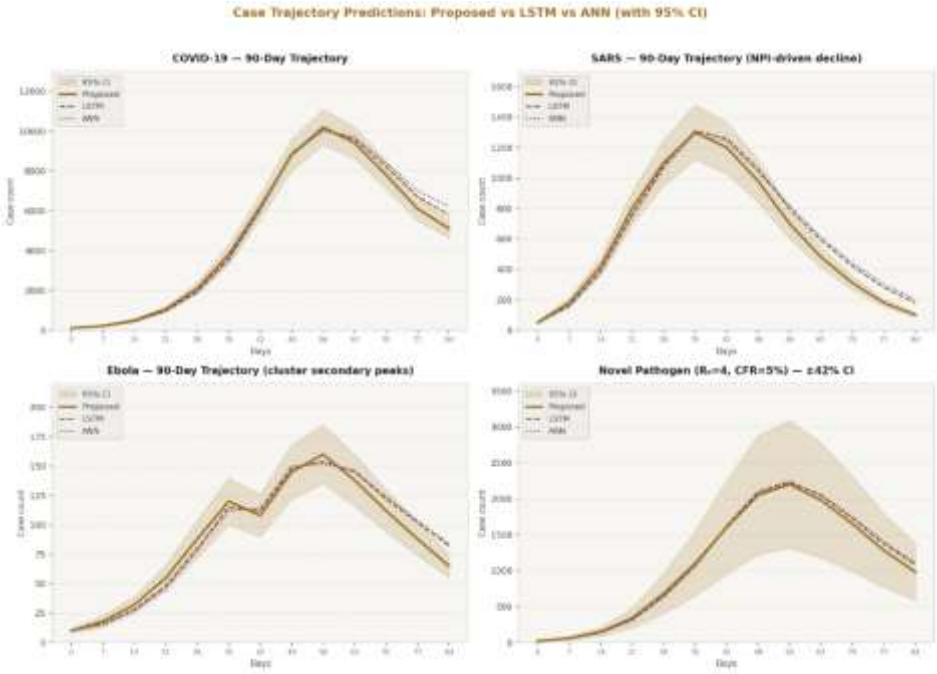

Figure 5: Case trajectory prediction

As shown in the multidimensional capability radar in Figure 6. It compares the proposed model with standard models: SVM, Adaboost (Decision tree), KNN, LSTM and ANN. The radar chart makes the SVM gap visible: its R^2, novel-F1, and CI coverage dimensions are all zero (classifier only). Same with KNN and Adaboost models. Whereas the LSTM and ANN show better CI coverage by covering all six dimensions with scores around 0.6 to 0.7. The proposed model shows a better-covered hexagon across all six dimensions with a better score of 0.8. This concludes that the proposed hybrid model is a better performer among all the standard models in terms of accuracy, prediction, minimum error, and better fit. It also acts as a better classifier not only for the known spread but for the new pathogen as well.

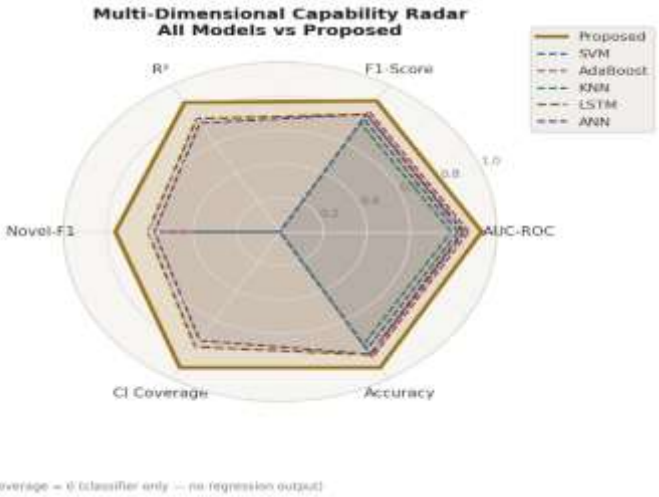

Figure 6: Multidimensional Capability Radar

5. CONCLUSION

This study presents a hybrid model that predicts outbreaks based on a three-layer machine-learning architecture. Including six pandemic datasets Spanish Flu, SARS, Zika, Swine Flu, Ebola and Covid-19, having 22 features like Virus Name, Infection Symptoms, Geographical Region, Number of infected cases (region wise), Number of Death (region wise), Time line, Case Fatality Rate (CFR), Basic Reproduction Number (R_0), Pandemic Duration, Global Mortality Estimates, Prevention, Control Measures and many. The first layer involves collecting and thoroughly pre-processing epidemic data. During pre-processing, an extensive range of techniques are applied to clean and normalise the data, including min-max normalisation, the generation of balanced datasets through data augmentation using SMOTE, and the retrieval of relevant features from the cleaned data. In order to identify the appropriate features to aid in epidemic prediction, several methods were applied to the pre-processed data, including central tendency, statistical dispersion, and time-series analysis, such as autocorrelation. A dimensionality reduction model, WMPCA, was applied for selecting the highly weighted features and optimal supporting features for further processing. Finally, a three-layered approach, a robust machine-learning model, was created to predict outbreaks. The first layer of the hybrid model is an AdaBoost SVM model for providing outbreak

probability and severity level; the second layer is KNN for providing trajectory class and spread pattern match for similar spreads; and the third layer uses an optimised ANN as the final predictor for predicted trajectory cases in 30/60/90 days. The weight of the O-ANN was optimised using a hybrid methodology, FAGWHP, which combines the strengths of FA and GWO to enhance prediction accuracy, which is 97.73. The proposed model shows the best evaluation score in terms of accuracy, RMSE, F1, R^2, and AUC-ROC with respect to the other standard models like SVM, ANN, LSTM, KNN and Adaboost.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used [Gemini, Chat GPT, and Google AI / scientific writing] to [better representation of work]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFERENCE

[1] X. X. Liu, S. J. Fong, N. Dey, R. G. Crespo, and E. Herrera-Viedma, "A new SEAIRD pandemic prediction model with clinical and epidemiological data analysis on COVID-19 outbreak," Applied Intelligence, vol. 51, pp. 4162–4198, 2021.

[2] S. F. Ardabili et al., "Covid-19 outbreak prediction with machine learning," Algorithms, vol. 13, no. 10, p. 249, 2020.

[3] C. Pizzuti, A. Socievole, B. Prasse, and P. Van Mieghem, "Network-based prediction of COVID-19 epidemic spreading in Italy," Appl. Netw. Sci., vol. 5, pp. 1–22, 2020.

[4] D. P. Kavadi, R. Patan, M. Ramachandran, and A. H. Gandomi, "Partial derivative nonlinear global pandemic machine learning prediction of covid 19," Chaos Solitons Fractals, vol. 139, p. 110056, 2020.

[5] A. Heidari, N. Jafari Navimipour, M. Unal, and S. Toumaj, "Machine learning applications for COVID-19 outbreak management," Neural Comput. Appl., vol. 34, no. 18, pp. 15313–15348, 2022.

[6] A. Khakharia et al., "Outbreak prediction of COVID-19 for dense and populated countries using machine learning," Annals of Data Science, vol. 8, pp. 1–19, 2021.

[7] A. E. Azzaoui, S. K. Singh, and J. H. Park, "SNS big data analysis framework for COVID-19 outbreak prediction in smart healthy city," Sustain. Cities Soc., vol. 71, p. 102993, 2021.

[8] S. Raheja, S. Kasturia, X. Cheng, and M. Kumar, "Machine learning-based diffusion model for prediction of coronavirus-19 outbreak," Neural Comput. Appl., vol. 35, no. 19, pp. 13755–13774, 2023.

[9] I. Ahmed, M. Ahmad, G. Jeon, and F. Piccialli, "A framework for pandemic prediction using big data analytics," Big Data Research, vol. 25, p. 100190, 2021.

[10] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi, and R. Gloaguen, "COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach," Mathematics, vol. 8, no. 6, p. 890, 2020.

[11] M. Saqib, "Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model," Applied Intelligence, vol. 51, no. 5, pp. 2703–2713, 2021.

[12] D. Prathumwan, K. Trachoo, and I. Chaiya, "Mathematical modeling for prediction dynamics of the coronavirus disease 2019 (COVID-19) pandemic, quarantine control measures," Symmetry (Basel)., vol. 12, no. 9, p. 1404, 2020.

[13] N. Jain et al., "Prediction modelling of COVID using machine learning methods from B-cell dataset," Results Phys., vol. 21, p. 103813, 2021.

[14] N. Absar, N. Uddin, M. U. Khandaker, and H. Ullah, "The efficacy of deep learning-based LSTM model in forecasting the outbreak of contagious diseases," Infect. Dis. Model., vol. 7, no. 1, pp. 170–183, 2022.

[15] R. Anandan, T. Nalini, S. Chiwhane, M. Shanmuganathan, and P. Radhakrishnan, COVID-19 outbreak data analysis and prediction. Measurement: Sensors, 25, p, 2023.

[16] L. R. Kolozsv?ri et al., "Predicting the epidemic curve of the coronavirus (SARS-CoV-2) disease (COVID-19) using artificial intelligence: An application on the first and second waves," Inform. Med. Unlocked, vol. 25, p. 100691, 2021.

[17] A. Hoque, A. Malek, and K. R. A. Zaman, "Data analysis and prediction of the COVID-19 outbreak in the first and second waves for top 5 affected countries in the world," Nonlinear Dyn., vol. 109, no. 1, pp. 77–90, 2022.

[18] S. Dash, C. Chakraborty, S. K. Giri, and S. K. Pani, "Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics," Pattern Recognit. Lett., vol. 151, pp. 69–75, 2021.

[19] N. Gamal, S. Ghoniemy, H. M. Faheem, and N. A. Seada, "Sentiment-Based Spatiotemporal Prediction Framework for Pandemic Outbreaks Awareness Using Social Networks Data Classification," IEEE Access, vol. 10, pp. 76434–76469, 2022, doi: 10.1109/ACCESS.2022.3192417.

[20] B. Jang, I. Kim, and J. W. Kim, "Long-Term Influenza Outbreak Forecast Using Time-Precedence Correlation of Web Data," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 5, pp. 2400–2412, May 2023, doi: 10.1109/TNNLS.2021.3106637.

[21] D. Gaglione and others, "Adaptive Bayesian Learning and Forecasting of Epidemic Evolution?Data Analysis of the COVID-19 Outbreak," IEEE Access, vol. 8, pp. 175244–175264, 2020, doi: 10.1109/ACCESS.2020.3019922.

[22] P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics," Chaos Solitons Fractals, vol. 139, p. 110058, 2020.

[23] Y.-C. Jin, Q. Cao, K.-N. Wang, Y. Zhou, Y.-P. Cao, and X.-Y. Wang, "Prediction of COVID-19 Data Using Improved ARIMA-LSTM Hybrid Forecast Models," IEEE Access, vol. 11, pp. 67956–67967, 2023, doi: 10.1109/ACCESS.2023.3291999.

[24] D. Vekaria, A. Kumari, S. Tanwar, and N. Kumar, ?“boost: An AI-Based Data Analytics Scheme for COVID-19 Prediction and Economy Boosting,” IEEE Internet Things J., vol. 8, no. 21, pp. 15977–15989, Nov. 2021, doi: 10.1109/JIOT.2020.3047539.
[25] R. Niu, E. W. M. Wong, Y.-C. Chan, M. A. Van Wyk, and G. Chen, “Modeling the COVID-19 Pandemic Using an SEIHR Model with Human Migration,” IEEE Access, vol. 8, pp. 195503–195514, 2020, doi: 10.1109/ACCESS.2020.3032584.
[26] S. Dhargupta, M. Ghosh, S. Mirjalili, and R. Sarkar, “Selective opposition based grey wolf optimization,” Expert Syst. Appl., vol. 151, p. 113389, 2020.
[27] K. Rahul and R. K. Banyal, “Firefly algorithm: an optimization solution in big data processing for the healthcare and engineering sector,” Int. J. Speech Technol., vol. 24, pp. 581–592, 2021.
[28] Manuela Rozalia Gabor, “A ‘new’ non-parametrical statistics instruments: Friedman test. Theoretical considerations and particularities for marketing data,” in International Day in Statistics & Economics, Prague, Sep. 2012.
[29] D. Rey and M. Neuhäuser, “Wilcoxon-Signed-Rank Test,” in International Encyclopedia of Statistical Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1658–1659. doi: 10.1007/978-3-642-04898-2_616.
[30] Kris N. Kirby and Daniel Gerlanc, “Finding Bootstrap Confidence Intervals for Effect Sizes With BootES,” APS Homepage.
[31] F. X. Diebold and R. S. Mariano, “Comparing Predictive Accuracy,” Journal of Business & Economic Statistics, vol. 13, no. 3, pp. 253–263, Jul. 1995, doi: 10.1080/07350015.1995.10524599.
[32] Z. Fan, E. Liu, and B. Xu, “Weighted Principal Component Analysis,” 2011, pp. 569–574. doi: 10.1007/978-3-642-23896-3_70.
[33] J. Shuja, E. Alanazi, W. Alasmary, and A. Alashaikh, “COVID-19 open source data sets: a comprehensive survey,” Applied Intelligence, vol. 51, no. 3, pp. 1296–1325, Mar. 2021, doi: 10.1007/s10489-020-01862-6.
[34] C. T. Yang et al., “Influenza-like illness prediction using a long short-term memory deep learning model with multiple open data sources,” J. Supercomput., vol. 76, pp. 9303–9329, 2020.
[35] P. D. Singh, R. Kaur, K. D. Singh, G. Dhiman, and M. Soni, “Fog-centric IoT based smart healthcare support service for monitoring and controlling an epidemic of Swine Flu virus,” Inform. Med. Unlocked, vol. 26, p. 100636, 2021.
[36] A. Erkoreka, “The Spanish influenza pandemic in occidental Europe (1918–1920) and victim age,” Influenza Other Respir. Viruses, vol. 4, no. 2, pp. 81–89, Mar. 2010, doi: 10.1111/j.1750-2659.2009.00125.x.
[37] Y. Muhammad, M. D. Alshehri, W. M. Alenazy, T. Vinh Hoang, and R. Alturki, “Identification of pneumonia disease applying an intelligent computational framework based on deep learning and machine learning techniques,” Mobile Information Systems, vol. 2021, pp. 1–20, 2021.
[38] M. Roser, “The Spanish flu: the global impact of the largest influenza pandemic in history,” Our World in Data, 2020.
[39] J. Zhang et al., “Viral pneumonia screening on chest X-rays using confidence-aware anomaly detection,” IEEE Trans. Med. Imaging, vol. 40, no. 3, pp. 879–890, 2020.
[40] A. Kara, “Multi-step influenza outbreak forecasting using deep LSTM network and genetic algorithm,” Expert Syst. Appl., vol. 180, p. 115153, 2021.