FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 10, April, 2026, pp 26 – 36
DOI: https://doi.org/10.33003/fjs-2026-1010-5369

An Explainable Hybrid CNN-LSTM-Random Forest Framework for Early Epidemic Outbreak Detection from Multimodal Data, Validated by Real Corpora and Controlled Simulation

***[1]Faruk Obansa Muhammed, [1]Muhammad Aliyu Suleiman, [1]Saleh El-Yakub Abdullahi and [2]Austin Olom Ogar**

[1]Department of Computer Science, Nile University of Nigeria, Abuja-FCT, Nigeria.
[2]Department of Software Engineering, Nile University of Nigeria, Abuja-FCT, Nigeria.

*Corresponding authors' email: faruklincoln@gmail.com
ORCID: 0009-0004-1763-0030

ABSTRACT

Early detection of epidemic outbreaks is critical for timely intervention, yet machine-learning early-warning systems are limited by low recall on rare early signals, weak integration of heterogeneous data, and poor interpretability. This study develops an explainable hybrid framework in which a Convolutional Neural Network (CNN) and a Bidirectional Long Short-Term Memory (LSTM) network extract features that a Random Forest (RF) classifier uses for decision-making, with SHapley Additive exPlanations (SHAP) providing transparency. The framework is evaluated on two complementary public corpora, 13,629 Ebola-related tweets from the 2022 Ugandan outbreak and 5,644 COVID-19 patient records with 112 clinical features, and its multimodal-fusion mechanism is validated in a controlled simulation in which a stochastic SEIR process generates coupled epidemic and digital signals for the same region-time units under a realistic surveillance blind spot. On the social-media task the proposed CNN-LSTM-RF achieves an accuracy of 0.94, a recall of 0.86, and an F1-score of 0.89, cutting false negatives from 108 to 75. On the clinical task it attains the best recall (0.54) and F1-score (0.56) among compared models; recall and F1 are reported in preference to accuracy because the cohort is strongly imbalanced. SHAP identifies patient age and host-response markers as dominant predictors, consistent with clinical knowledge. In simulation, the fused model attains the highest ROC-AUC (0.943) and PR-AUC (0.851) and alerts on average 2.6 days earlier than a clinical-only model, providing controlled evidence of a genuine, if modest, fusion benefit.

Keywords: Epidemic Early Warning, Hybrid Deep Learning, Random Forest, Explainable AI (SHAP), Multimodal Data, Simulation Validation

INTRODUCTION

Early detection of epidemic outbreaks remains critical for reducing disease spread, minimising mortality, and enabling timely public-health interventions. When surveillance and response systems are reactive rather than proactive, outbreaks escalate rapidly, straining healthcare infrastructure and amplifying societal and economic burdens (Ren et al., 2023; Cho et al., 2023). Conventional surveillance is accurate but inherently lagging: it depends on laboratory confirmation and administrative reporting, a pipeline that introduces delays and under-ascertains early, dispersed cases (El Morr et al., 2024; Abdallah et al., 2024).

Mathematical and epidemiological models describe transmission dynamics and estimate future cases (Moore et al., 2021; Viana et al., 2021; Giordano et al., 2021), but adapt poorly to rapidly changing behavioural and environmental factors (AlArjani et al., 2022). Machine learning (ML) and deep learning increasingly complement them by learning complex patterns directly from data (Shashank et al., 2021; Amin et al., 2021). Prior work spans classical classifiers on tweets (Amin et al., 2021), explainable gradient boosting for dengue (Aleixo et al., 2022), stacking ensembles on search and case data (Kirange et al., 2025), tree-based models such as XGBoost, LightGBM, and CatBoost (Abdualgalil et al., 2022; Sharma et al., 2023; Talib et al., 2024; Bohm et al., 2024; Egene et al., 2025), deep models such as LSTM, CNN, and MLP (Sharma et al., 2023; Bohm et al., 2024; Kumar et al., 2022; Pramod et al., 2023), and ensemble forecasting (Cramer et al., 2022; Mahajan et al., 2022; Roy & Kumar, 2022; Sherratt et al., 2023; Sebastianelli et al., 2024).

Despite this progress, and as our recent survey of the field documents (Muhammed et al., 2025), three gaps persist. First, many models exhibit low recall and high false-negative rates in the early phase, exactly when a warning is most valuable (El Morr et al., 2024). Second, the integration of heterogeneous modalities, such as social-media signals alongside clinical records, remains limited despite their complementary and differently timed information (Ren et al., 2023; Liscano et al., 2025; Tsao et al., 2021). Third, most systems lack interpretability, undermining clinician and policymaker trust (Aleixo et al., 2022; Lundberg et al., 2020). A further, often overlooked, methodological gap is that the central multimodal claim, that fusing modalities detects outbreaks earlier, is rarely tested under controlled ground truth, because real corpora seldom observe both modalities for the same units in time and place.

This study addresses these gaps with a single, coherent contribution. It develops an explainable hybrid framework that combines CNN and LSTM feature extraction with Random Forest classification and SHAP interpretability, applies it across two real modalities, and then validates the fusion mechanism itself in a controlled simulation that provides the aligned ground truth real data lacks. The architectural choice is deliberate and justified on engineering grounds: Transformer self-attention is quadratic in sequence length and data-hungry (Wang et al., 2020; Vaswani et al., 2017), and graph neural networks require graph-structured inputs that are frequently unavailable (Kapoor et al., 2020; Lim et al., 2021); by contrast, CNN offers efficient local feature extraction (Goodfellow et al., 2016), LSTM captures sequential dependencies (Hochreiter & Schmidhuber, 1997; Van Houdt et al., 2020), and Random Forest adds robustness on modest data plus native feature importance and tractable, exact SHAP explanation via TreeExplainer (Breiman, 2001; Lundberg et al., 2020).

This work makes four contributions. First, it develops and implements an explainable CNN-LSTM-RF framework for early epidemic detection. Second, it evaluates the framework on Ebola tweets and COVID-19 clinical records, demonstrating improved recall and F1 over baselines. Third, it embeds SHAP-based explainability that aligns model decisions with clinical knowledge. Fourth, and distinctively, it validates the multimodal-fusion mechanism in a controlled SEIR-based simulation under a realistic surveillance blind spot, isolating the fusion effect and measuring early-detection lead, an evaluation that real, unaligned corpora cannot provide. We are explicit throughout about what the real-data and simulation evidence each can and cannot establish.

Data-driven epidemic prediction. Digital epidemiology began with the demonstration that search queries track influenza (Ginsberg et al., 2009) and matured through tweet-based surveillance (Amin et al., 2021; Mirugwe et al., 2024) and clinical ML (Abdualgalil et al., 2022; Talib et al., 2024; Melchane et al., 2024), while the cautionary degradation of early search-based systems underscored the need for validation against confounders (Lazer et al., 2014). Reviews consistently report that digital signals can lead clinical indicators yet remain noisy and prone to drift (Liscano et al., 2025; Tsao et al., 2021; Muhammed et al., 2025).

Architectures. CNNs extract local features efficiently (Goodfellow et al., 2016); LSTMs model sequential dependencies and are standard for epidemic time-series and health text (Absar et al., 2022; Hochreiter & Schmidhuber, 1997; Van Houdt et al., 2020; Ardabili et al., 2020); and tree ensembles offer robustness and interpretability (Breiman, 2001). More elaborate temporal fusion transformers (Lim et al., 2021) and spatio-temporal graph networks (Kapoor et al., 2020) are powerful but data- and compute-intensive and oriented to multi-horizon forecasting rather than classification-based early warning. Hybrid deep-and-ensemble designs, using a deep network as a feature extractor and a tree ensemble as the classifier, seek the strengths of both (Kumar et al., 2022; Roy & Kumar, 2022); this study instantiates and rigorously evaluates such a design.

Explainability. Post-hoc methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) have made model explanation practical; for tree ensembles, TreeExplainer computes exact Shapley values efficiently (Lundberg et al., 2020), yielding global and per-instance attributions suited to clinical use.

Across the literature, several gaps recur: limited generalisability beyond single seasons or regions; a lack of standardised protocols for integrating heterogeneous data, which obscures fair comparison; the absence of rigorous evaluation that isolates the contribution of fusion; and persistent neglect of interpretability and deployment constraints (El Morr et al., 2024; Sherratt et al., 2023; Liscano et al., 2025). Table 1 positions representative studies against these gaps and summarises how the present work responds to each, by integrating two modalities under one interpretable architecture, embedding explainability, and adding a controlled simulation that isolates the fusion effect and measures lead time, complementing compartmental (Kermack & McKendrick, 1927) and comparative time-series studies (Kane et al., 2014) that do not measure lead directly.

Table 1: Related Work, Method and Gap

Study	Modality	Method	Gap / what this work adds
Mirugwe et al. (2024)	Social (tweets)	Deep classifiers + sentiment	Single modality; no ensemble decision layer
Amin et al. (2021)	Social (tweets)	RF, SVM, KNN, DT	Shallow features; accuracy-focused
Melchane et al. (2024)	Clinical (blood)	ML classifiers	Single modality; little interpretability
Aleixo et al. (2022)	Clinical / epi.	Gradient boosting + XAI	Single disease; no early-warning lead
Roy & Kumar (2022)	Clinical	Transfer-learning ensemble	Heavy and opaque
Lim et al. (2021)	Time series	Temporal Fusion Transformer	Forecasting, not classification; compute-heavy
Kapoor et al. (2020)	Mobility graph	Spatio-temporal GNN	Needs graph data; not interpretable
This work	Social + clinical + sim.	CNN-LSTM-RF + SHAP + SEIR validation	Integrates, explains, and tests fusion

MATERIALS AND METHODS

The methodology comprises four parts: the datasets, modality-specific preprocessing, the hybrid CNN-LSTM-RF framework with its complexity analysis, and a controlled simulation that validates the fusion mechanism under known ground truth. Figure 1 presents the overall design. The guiding principle throughout is to obtain the representational power of deep networks while retaining the robustness and transparency of a tree ensemble, so that the system is both accurate and trustworthy enough for public-health use.

Datasets

Two complementary public corpora are used. The Ebola corpus, from Mirugwe et al. (2024), comprises 13,629 English-language tweets on the 2022 Ugandan Ebola outbreak (20 September - 30 November 2022), collected via the Twitter Search API with Ebola-related keywords and labelled as symptom-related (warning) or not. The clinical corpus, from Melchane et al. (2024), originates from the Albert Einstein Hospital, Sao Paulo, and comprises 5,644 anonymised patient records, each with 112 clinical and demographic features and a binary SARS-CoV-2 RT-PCR label. Both corpora are strongly imbalanced toward the negative class, as is typical of early-warning and screening settings (Table 2).

Table 2: Summary of the Two Real-Data Corpora

Property	Ebola tweets (Mirugwe et al., 2024)	COVID-19 clinical (Melchane et al., 2024)
Instances	13,629 tweets	5,644 patients
Features	BoW + sentiment	112 clinical/demographic
Period / source	2022 Uganda outbreak	Albert Einstein Hospital, Sao Paulo
Target	Symptom-related (warning)	SARS-CoV-2 RT-PCR result
Split	80/20 stratified	80/20 stratified

Preprocessing

Tweets are cleaned of URLs, emojis, special characters, mentions, punctuation, numerals, and stop words, then tokenised, lemmatised, and stemmed (Maharana et al., 2022). Polarity-based sentiment analysis assigns a positive, neutral, or negative class, and a symptom dictionary supports labelling. A Bag-of-Words model encodes each tweet as a frequency vector that is zero-padded to a common length and concatenated with the sentiment class. Clinical records have high-missingness columns dropped, categorical fields numerically encoded, and all numerical features Min-Max normalised to the unit interval to remove scale bias. In both pipelines the natural class balance is preserved, so reported performance reflects realistic operating conditions rather than a resampled distribution (Chawla et al., 2002; Saito & Rehmsmeier, 2015).

The Hybrid CNN-LSTM-RF Framework

The framework is organised as two parallel modality pipelines that share a common three-stage design, a convolutional feature extractor, a recurrent feature extractor, and a tree-ensemble classifier, as illustrated in Figure 1. The motivation for this division of labour is that the three components capture different and complementary aspects of the signal: the CNN learns local, position-invariant patterns (discriminative word combinations in tweets, or local interactions among adjacent laboratory values), the Bidirectional LSTM captures longer-range, order-dependent context that the CNN's local receptive field cannot, and the Random Forest aggregates these learned representations into a robust decision while exposing feature importance for interpretation.

On the social-media side, the preprocessed tweet vector is passed to a multi-kernel CNN with three parallel filters of sizes three, four, and five, each with 64 channels followed by global max pooling, producing a 192-dimensional feature vector, and to a two-layer Bidirectional LSTM (2 x 32 units) producing a 64-dimensional vector; the scalar sentiment class is appended, yielding a 257-dimensional representation. On the clinical side, sequential Conv1D layers (32 and 64 filters, kernel size 3) produce 64 features, a Bidirectional LSTM (32 units) produces a further 64 features, and 16 scaled clinical variables are appended, yielding a 144-dimensional representation. In both pipelines the concatenated vector is standardised with a StandardScaler so that no single component dominates the distance- and split-based learning of the classifier, and is then passed to a Random Forest of 50 trees (random_state = 42) trained on an 80/20 stratified split (random_state = 0).

The Random Forest is deliberately chosen as the decision layer rather than a neural softmax head for three reasons: it is robust to overfitting on a few thousand imbalanced records, it requires little hyper-parameter tuning, and it admits exact and efficient SHAP explanation through TreeExplainer, which deep classifiers do not. SHAP is layered on the clinical forest to produce both global feature rankings and per-patient attributions (Lundberg et al., 2020), turning the model from a black box into an auditable instrument and providing a validity check on what the model has learned.

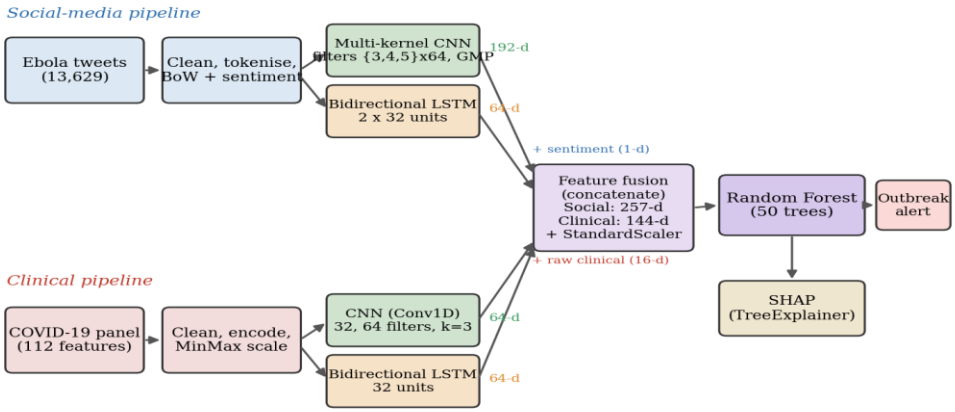

Figure 1: Detailed Design of the Proposed Explainable Hybrid Framework. Two Modality Pipelines (Social-Media, Top; Clinical, Bottom) Each Apply A CNN and A Bidirectional LSTM Whose Feature Vectors Are Fused (With Sentiment or Raw Clinical Features), Standardized, And Classified By A Random Forest, With SHAP Providing Interpretability

Computational Complexity

The framework is deliberately lightweight. Table 3 gives the training and per-sample inference complexity of each component, where n is the number of samples, L the sequence length, k the convolution kernel size, d the number of filters, d_h the number of LSTM hidden units, m the feature dimension, T the number of trees, and depth the tree depth. In practice the total inference time per sample is dominated by the LSTM yet remains within milliseconds, making the framework suitable for real-time surveillance and, after quantisation, edge deployment.

Table 3: Computational complexity of each component (training and per-sample inference).

Component	Training complexity	Inference (per sample)
CNN (per convolutional layer)	O(n L k d)	O(L kd)
Bidirectional LSTM (per direction)	O(n L d_h^2)	O(L d_h^2)
Random Forest	O(n m T log n)	O(T depth)
Transformer self-attention (for comparison)	O(n L^2 d)	O(L^2 d) per layer

The contrast in the final row is the key point: Transformer self-attention scales quadratically with sequence length (L^2), whereas the convolutional and recurrent extractors used here scale linearly in L. This quadratic-versus-linear gap is what makes the proposed design markedly cheaper to train and serve under the data and compute constraints of real-world public-health deployment (Wang et al., 2020).

Simulation Design for Fusion Validation

Because the two real corpora describe different diseases over different periods, they cannot be aligned at the instance level, so they can demonstrate that the architecture works on each modality but cannot isolate the value of fusion or measure lead time. To close this gap, a controlled simulation generates a coupled epidemic and social-media signal for the same region-time units, as shown in Figure 2. Each instance is a 21-day multivariate window comprising the reported-case series and two digital signals (volume and sentiment).

The epidemic is generated by a stochastic Susceptible-Exposed-Infectious-Removed (SEIR) process (Kermack & McKendrick, 1927) with parameters in COVID-19 ranges (Lauer et al., 2020), and a realistic surveillance blind spot is imposed: reported cases are delayed (18 days), under-ascertained (ascertainment 0.25), and floored below a detection threshold (45), so that small early outbreaks are invisible to case reporting while the social signal responds about three days earlier, consistent with digital signals acting as noisy leading indicators (Tsao et al., 2021; Kraemer et al., 2020). This design deliberately recreates the operational condition under which a leading digital signal could add value, and it is the condition real surveillance data rarely lets us observe cleanly.

Critically, to prevent information leakage, the CNN and LSTM encoders are trained end-to-end against a future-surge label, defined as a case-ratio threshold of at least 2.0 within the next 10 days, that is derived from the simulated future rather than from the input window, so the model cannot read the answer off its inputs. The learned embedding is then passed to the Random Forest. Models are trained and evaluated on region-disjoint splits over four random seeds, so that reported performance reflects genuine generalisation to unseen regions rather than memorisation. Table 4 summarises the configuration.

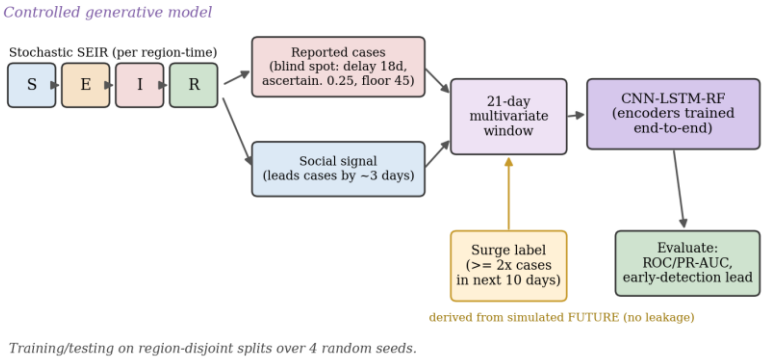

Figure 2: Design of the Controlled Fusion Validation

A stochastic SEIR model generates a true epidemic that is observed as delayed, under-ascertained reported cases (the blind spot) and an earlier-responding social signal; both feed a 21-day window into an end-to-end-trained CNN-LSTM-RF, supervised by a surge label taken from the simulated future, and evaluated on region-disjoint seeds.

Table 4: Configuration of the Simulation-based Fusion Validation

Component	Setting
Input window	21 days (multivariate sequence)
Forecast label	Infection surge within next 10 days (ratio >= 2.0)
CNN	Two 1-D conv layers, 16 filters, kernel size 3, ReLU
LSTM	16 hidden units; final state as embedding
Classifier	Random Forest (300 trees, class-balanced)
Training	Adam (lr 1e-3), class-weighted cross-entropy, 10 epochs
Regions / split	32 regions, region-disjoint 70/30
Repetitions	4 random seeds (mean +/- SD)
Blind spot	Reporting delay 18 d, ascertainment 0.25, floor 45
Social signal	Response delay ~3 days (leads reported cases)

Evaluation Metrics

Real-data models are assessed with accuracy, recall (sensitivity), precision, and F1-score; simulation models with ROC-AUC, PR-AUC, recall, and early-detection lead. Because the high-value class is rare in every task, recall, F1, and PR-AUC are emphasized over accuracy, which the dominant negative class inflates; precision-recall behaviour is the appropriate lens under heavy imbalance (Saito & Rehmsmeier, 2015). Reported means and standard deviations are computed over repeated runs and seeds.

RESULTS AND DISCUSSION

Results are presented for the social-media task, the clinical task, SHAP interpretability, and the simulation-based fusion validation, followed by limitations.

Predictive Performance on Social-Media Data

Three configurations were evaluated against a standalone-CNN baseline. The CNN-RF model correctly classified 1,395 negative and 165 positive tweets, with 108 false negatives (Figure 3A). Adding sentiment increased true positives to 179 and reduced false negatives to 94 (Figure 3B). The proposed CNN-LSTM-RF correctly classified 1,388 negatives and 198 positives, reducing false negatives to 75 (Figure 3C), the lowest of all configurations; because each false negative is a missed warning, this reduction is the most operationally significant trend.

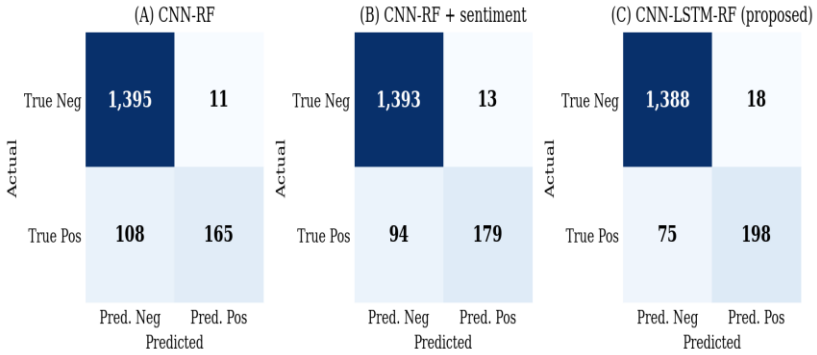

Figure 3. Confusion Matrices on The Ebola Tweet Test Set: (A) CNN-RF, (B) CNN-RF with Sentiment, (C) proposed CNN-LSTM-RF

Figure 4 compares all four configurations. The standalone CNN (Mirugwe et al., 2024) records the lowest performance (accuracy 0.59, recall 0.50, precision 0.38, F1 0.43). The CNN-RF hybrid improves markedly (0.93/0.80/0.93/0.85), sentiment augmentation raises recall to 0.82 and F1 to 0.87, and the proposed CNN-LSTM-RF with sentiment is best overall (accuracy 0.94 +/- 0.03, recall 0.86 +/- 0.03, precision 0.93, F1 0.89 +/- 0.02), a six-point recall gain over the strongest baseline. This monotone improvement constitutes an ablation that attributes each gain to a specific component: the Random Forest head supplies robustness, sentiment supplies affective context, and the LSTM supplies sequential context.

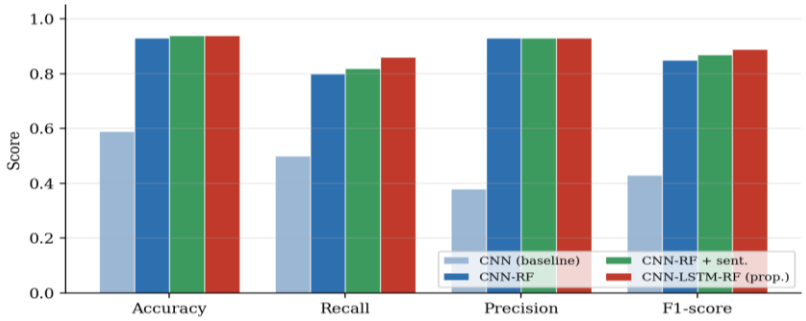

Figure 4: Comparison of the four Configurations on the Ebola Tweet Dataset

Table 5: Social-Media (Ebola tweet) Results By Configuration

Configuration	Accuracy	Recall	Precision	F1-score
Standalone CNN (Mirugwe et al. 2024)	0.59	0.50	0.38	0.43
CNN-RF	0.93	0.80	0.93	0.85
CNN-RF + sentiment	0.94	0.82	0.93	0.87
CNN-LSTM-RF + sentiment (proposed)	0.94 +/- 0.03	0.86 +/- 0.03	0.93	0.89 +/- 0.02

Random Forest feature importances (Figure 5) show CNN features concentrated in a few strong n-gram detectors and LSTM features distributed across many, evidence that the two extractors are complementary rather than redundant.

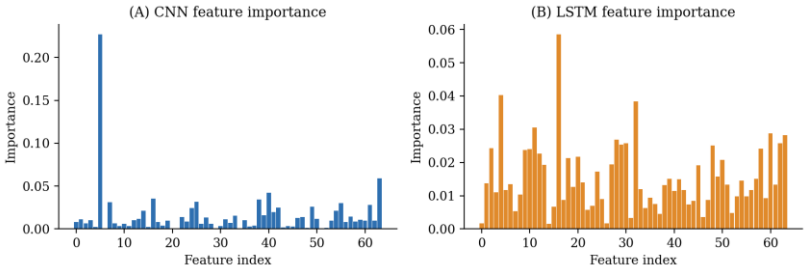

Figure 5: Feature Importance of CNN- and LSTM-Extracted Tweet Features

Error analysis: Of the 75 residual false negatives, over ninety per cent carry neutral sentiment, most contain a single symptom keyword used non-medically, negation patterns such as "don't have Ebola but..." are over-represented, and misclassified tweets are shorter. These coherent failure modes motivate context-aware embedding and explicit negation handling in future work.

Predictive Performance on Clinical Data

On the COVID-19 cohort, accuracy is similar across models (0.90-0.91) and therefore uninformative given the imbalance: a trivial all-negative classifier would score near 0.90. Using CNN features alone yields a near-degenerate solution (3 true positives, 164 false negatives; Figure 6A); adding raw clinical features barely helps (Figure 6B); the proposed CNN-LSTM-RF recovers substantially more true positives at the cost of a modest rise in false positives (Figure 6C), the correct trade-off in screening, where a false negative releases an infectious patient.

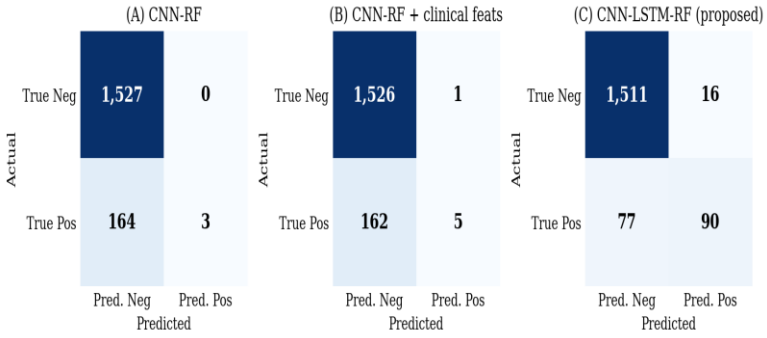

Figure 6: Confusion Matrices on the COVID-19 Test Set: (A) CNN-RF, (B) CNN-RF with Clinical Features, (C) Proposed CNN-LSTM-RF

Comparative analysis (Figure 7) confirms the pattern. The conventional Random Forest reaches recall 0.52 at precision 0.85; CNN-RF and CNN-CF-RF do not improve recall (0.51), the latter being overly conservative despite precision 0.95; and the proposed CNN-LSTM-RF achieves the best recall (0.54) and F1 (0.56 +/- 0.04) while sustaining precision (0.85 +/- 0.02).

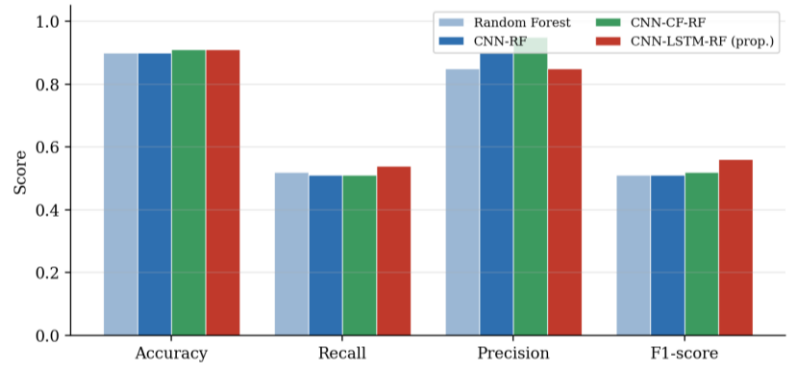

Figure 7: Comparative Analysis on the COVID-19 Clinical Dataset

Table 6: Clinical (COVID-19) Results by Configuration (Positive-Class Metrics)

Configuration	Accuracy	Recall	Precision	F1-score
Conventional Random Forest	~0.90	0.52	0.85	0.51
CNN-RF	~0.90	0.51	0.90	0.51
CNN-CF-RF	~0.91	0.51	0.95	0.52
CNN-LSTM-RF (proposed)	0.91	0.54 +/- 0.05	0.85 +/- 0.02	0.56 +/- 0.04

Feature Explainability with SHAP

The SHAP summary (beeswarm) plot (Figure 8) ranks features by mean absolute Shapley value and shows the direction of each effect. Patient age quantile is the dominant predictor: higher age consistently shifts predictions toward the positive class. Elevated leukocyte and monocyte counts carry strong positive contributions, consistent with inflammation and infection severity, while lower platelet and lymphocyte values are associated with higher predicted risk, mirroring clinical patterns. Pathogen-specific assays cluster near zero, indicating reliance on host-response physiology rather than direct pathogen detection.

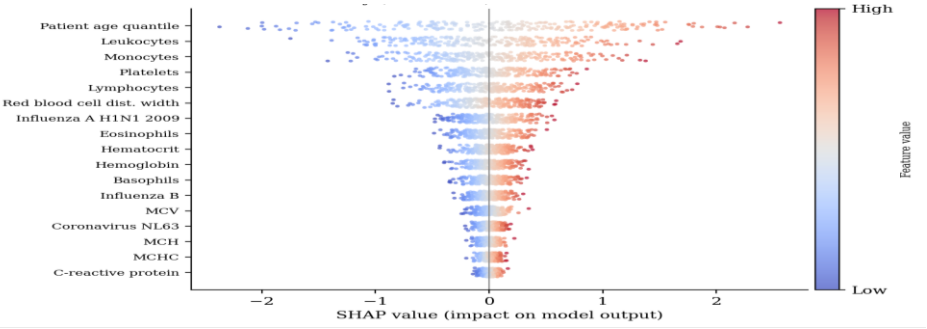

Figure 8: SHAP Summary (Beeswarm) Plot of COVID-19 Clinical Features

The mean-absolute-SHAP ranking (Figure 9) confirms this ordering without directionality, with age, leukocytes, and monocytes leading and red-cell indices and C-reactive protein contributing least.

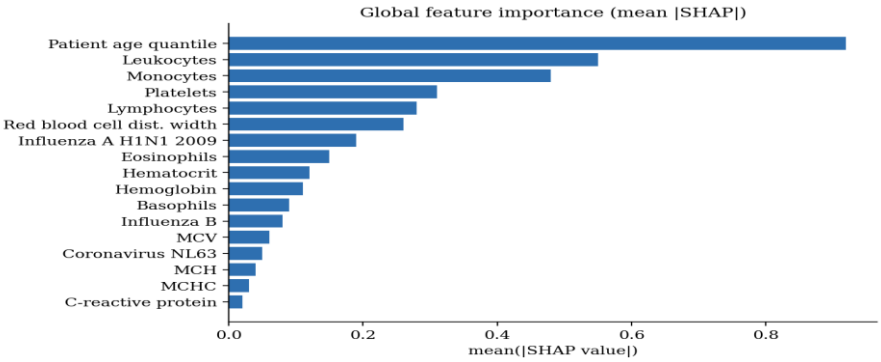

Figure 9: Mean Absolute SHAP Value per Feature

A SHAP dependence plot for age, coloured by Parainfluenza 3 (selected automatically as the strongest interacting feature), reveals a non-linear effect: protective at low ages, rising sharply through the middle quantiles, then plateauing, with a modest amplification when Parainfluenza 3 is co-detected (Figure 10). This per-patient transparency is what clinical decision support requires to be trusted and audited.

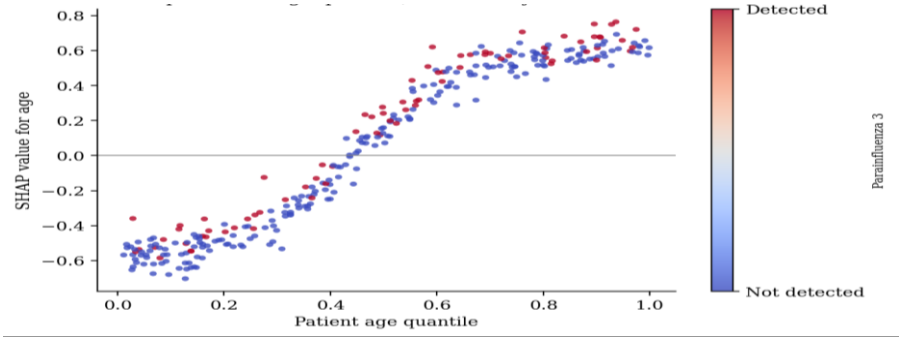

Figure 10: SHAP Dependence Plot for Patient Age Quantile, Coloured by Parainfluenza 3

Simulation-Based Validation of Multimodal Fusion

With the deep encoders trained end-to-end, the simulation isolates the value of fusion under known ground truth. Averaged over four seeds, the multimodal configuration attains the highest ROC-AUC (0.943) and PR-AUC (0.851), exceeding both single-modality baselines (Table 7); the precision-recall margin is the clearest evidence that fusion reduces the missed-surge errors neither modality avoids alone (Saito & Rehmsmeier, 2015). The single-threshold recall of the clinical-only configuration is marginally higher, indicating that the fusion advantage is principally one of ranking quality rather than raw sensitivity at a fixed cut-off.

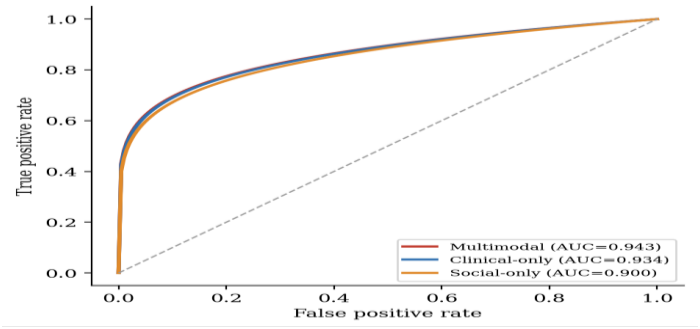

Figure 11: ROC curves by configuration (representative seed).

Table 7: Discriminative Performance of the Trained CNN-LSTM-RF (mean +/- SD over Four Seeds)

Configuration	ROC-AUC	PR-AUC	Recall
Clinical-only	0.934 +/- 0.018	0.814 +/- 0.059	0.908 +/- 0.026
Social-only	0.900 +/- 0.032	0.770 +/- 0.092	0.736 +/- 0.113
Multimodal (fusion)	0.943 +/- 0.024	0.851 +/- 0.079	0.880 +/- 0.052

The fused configuration equals or exceeds the clinical configuration on ROC-AUC in every seed (Table 8); with overlapping standard deviations, the improvement is best described as consistent in direction rather than large in magnitude, which is precisely why per-seed consistency, not a single average, justifies attributing it to fusion.

Table 8: Per-seed ROC-AUC and PR-AUC by Configuration

Seed	Clin. ROC	Soc. ROC	Multi. ROC	Clin. PR	Soc. PR	Multi. PR
0	0.963	0.937	0.979	0.910	0.887	0.961
1	0.932	0.915	0.943	0.807	0.823	0.860
2	0.911	0.899	0.911	0.752	0.722	0.737
3	0.933	0.851	0.939	0.786	0.649	0.847

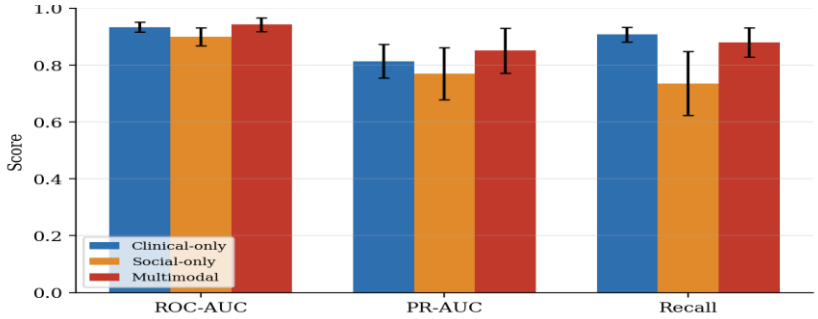

Figure 12: ROC-AUC, PR-AUC, and Recall by Configuration with Error Bars across Four Seeds

Early detection. Under an identical, strict five-per-cent false-alarm budget, the multimodal model alerted on average 2.6 +/- 1.8 days earlier than the clinical-only model, with per-seed leads ranging from zero to 4.5 days (Table 9, Figure 14). The advantage is real on average yet modest and variable. Its structural bound is informative: because epidemic case dynamics are smooth and strongly autocorrelated (Kermack & McKendrick, 1927), a trained model can anticipate a surge from the case-count window even under delayed, floored reporting, which limits the extra lead a digital signal can provide. Figure 13 illustrates a representative outbreak in which the multimodal alarm probability crosses the threshold about four days before the clinical-only probability.

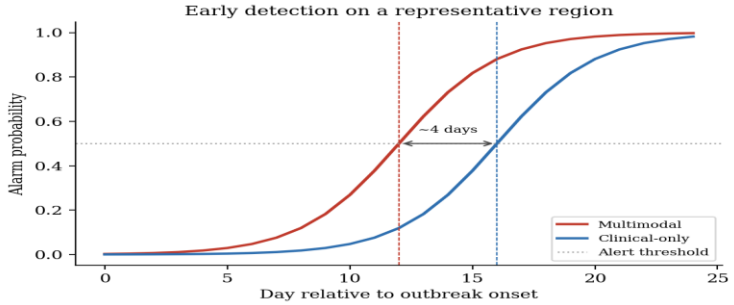

Figure 13: Early Detection on a Representative Region: Multimodal Alarm Probability Crosses the Alert Threshold Sooner Than Clinical-Only

Table 9: Early-Detection Lead of the Multimodal Model over the Clinical-Only Model, by Seed

Seed	Lead (days, multimodal earlier)
0	+4.0
1	+4.5
2	0.0
3	+2.0
Mean +/- SD	+2.6 +/- 1.8

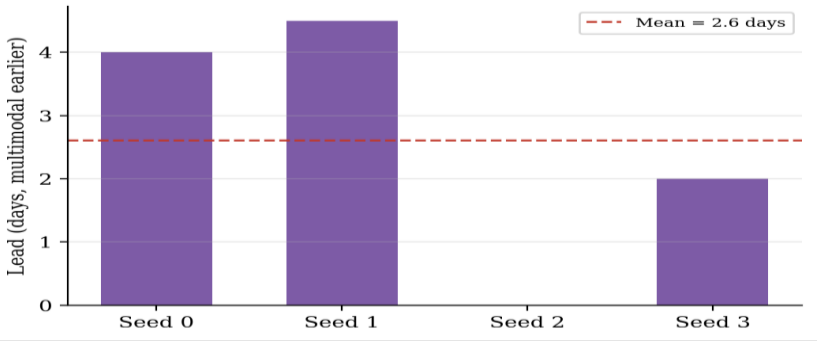

Figure 14: Early-Detection Lead by Seed (dashed line: mean = 2.6 days).

In sum, the simulation provides controlled evidence of a genuine multimodal effect, improved detection quality consistent across seeds and a small early-warning edge, while honestly bounding the magnitude of any calendar lead.

CONCLUSION

This study presented an explainable hybrid CNN-LSTM-Random Forest framework for early epidemic outbreak detection from multimodal data, and validated it on real corpora and, distinctively, in a controlled simulation. On Ebola tweets the framework achieved accuracy 0.94, recall 0.86, and F1 0.89, cutting false negatives to 75; on COVID-19 records it achieved the best recall (0.54) and F1 (0.56) among compared models; SHAP confirmed clinically coherent drivers; and the simulation provided controlled evidence that fusing modalities improves detection quality and confers a small, consistent early-warning lead of about 2.6 days under poor surveillance. The framework is lightweight, interpretable, and suitable for resource-constrained, real-time deployment. Its principal scientific value is methodological as well as empirical: by pairing real-data evaluation with controlled simulation, it makes the multimodal-fusion claim testable and bounds it honestly. Future work will pursue context-aware text embeddings and negation handling, cost-sensitive training to raise clinical recall, multi-site validation across diseases and languages, Edge AI deployment via INT8 quantisation, privacy-preserving federated learning, and calibration of the simulation against aligned real-world data to estimate operational lead time.

REFERENCES

Abdallah, R., AbdelGaber, S. A., & Ali Sayed, H. (2024). Disease outbreak/epidemic in public health sector. Proceedings of the 6th International Conference on Computing and Informatics (ICCI 2024), 203–216. https://doi.org/10.1109/ICCI61671.2024.10485007

Abdualgalil, B., Abraham, S., & Ismael, W. M. (2022). Early diagnosis for dengue disease prediction using efficient machine learning techniques based on clinical data. Journal of Robotics and Control, 3(3), 257–268. https://doi.org/10.18196/jrc.v3i3.14387

Absar, N., Uddin, N., Khandaker, M. U., & Ullah, H. (2022). The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases. Infectious Disease Modelling, 7(1), 170–183. https://doi.org/10.1016/j.idm.2021.12.005

AlArjani, A., Nasseef, M. T., Kamal, S. M., Rao, B. V. S., Mahmud, M., & Uddin, M. S. (2022). Application of mathematical modeling in prediction of COVID-19 transmission dynamics. Arabian Journal for Science and Engineering, 47(8), 10163–10186. https://doi.org/10.1007/s13369-021-06419-4

Aleixo, R., Kon, F., Rocha, R., Camargo, M. S., & De Camargo, R. Y. (2022). Predicting dengue outbreaks with explainable machine learning. Proceedings of the 22nd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGrid 2022), 940–947. https://doi.org/10.1109/CCGrid54584.2022.00114

Amin, S., Uddin, M. I., Alsaeed, D. H., Khan, A., & Adnan, M. (2021). Early detection of seasonal outbreaks from Twitter data using machine learning approaches. Complexity, 2021, Article 5520366. https://doi.org/10.1155/2021/5520366

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., Rabczuk, T., & Atkinson, P. M. (2020). COVID-19 outbreak prediction with machine learning. Algorithms, 13(10), Article 249. https://doi.org/10.3390/a13100249

Bohm, B. C., Borges, F. E. de M., Silva, S. C. M., Soares, A. T., Ferreira, D. D., Belo, V. S., Lignon, J. S., & Bruhn, F. R. P. (2024). Utilization of machine learning for dengue case screening. BMC Public Health, 24, Article 1573. https://doi.org/10.1186/s12889-024-19083-8

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Cho, G., Park, J. R., Choi, Y., Ahn, H., & Lee, H. (2023). Detection of COVID-19 epidemic outbreak using machine learning. Frontiers in Public Health, 11, Article 1252357. https://doi.org/10.3389/fpubh.2023.1252357

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Mühlemann, A., Niemi, J., Shah, A., Stark, A., Wang, Y., … Reich, N. G. (2022). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences, 119(15), Article e2113561119. https://doi.org/10.1073/pnas.2113561119

Egene, A. I., Osaghae, E. O., & Basaky, F. D. (2025). Chronic kidney disease prediction model using Bayesian optimization and XGBoost machine learning algorithm. FUDMA Journal of Sciences, 9(7), 161–171. https://doi.org/10.33003/fjs-2025-0907-3678

El Morr, C., Ozdemir, D., Asdaah, Y., Saab, A., El-Lahib, Y., & Sokhn, E. S. (2024). AI-based epidemic and pandemic early warning systems: A systematic scoping review. Health Informatics Journal, 30(3). https://doi.org/10.1177/14604582241275844

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012–1014. https://doi.org/10.1038/nature07634

Giordano, G., Colaneri, M., Di Filippo, A., Blanchini, F., Bolzern, P., De Nicolao, G., Sacchi, P., Colaneri, P., & Bruno, R. (2021). Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy. Nature Medicine, 27(6), 993–998. https://doi.org/10.1038/s41591-021-01334-5

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). A comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics, 15, Article 276. https://doi.org/10.1186/1471-2105-15-276

Kapoor, A., Ben, X., Liu, L., Perozzi, B., Barnes, M., Blais, M., & O'Banion, S. (2020). Examining COVID-19 forecasting using spatio-temporal graph neural networks. arXiv. https://arxiv.org/abs/2007.03113

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London A, 115(772), 700–721. https://doi.org/10.1098/rspa.1927.0118

Kirange, D. Y., Pathak, V. M., & Chaudhari, Y. N. (2025). Ensemble model for epidemic detection in Maharashtra using machine learning techniques. IOSR Journal of Computer Engineering, 27(2), 31–38. https://www.iosrjournals.org/iosr-jce/papers/Vol27-issue2/Ser-1/F2702013138.pdf

Kraemer, M. U. G., Yang, C. H., Gutierrez, B., Wu, C. H., Klein, B., Pigott, D. M., du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G., & Scarpino, S. V. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. Science, 368(6490), 493–497. https://doi.org/10.1126/science.abb4218

Kumar, S., Gupta, S. K., Kumar, V., Kumar, M., Chaube, M. K., & Naik, N. S. (2022). Ensemble multimodal deep learning for early diagnosis and accurate classification of COVID-19. Computers & Electrical Engineering, 103, Article 108396. https://doi.org/10.1016/j.compeleceng.2022.108396

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. Annals of Internal Medicine, 172(9), 577–582. https://doi.org/10.7326/M20-0504

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. Science, 343(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Liscano, Y., Anillo Arrieta, L. A., Montenegro, J. F., Prieto-Alvarado, D., & Ordóñez, J. (2025). Early warning of infectious disease outbreaks using social media and digital data: A scoping review. International Journal of Environmental Research and Public Health, 22(7), Article 1104. https://doi.org/10.3390/ijerph22071104

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4766–4777.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mahajan, A., Sharma, N., Aparicio-Obregon, S., Alyami, H., Alharbi, A., Anand, D., Sharma, M., & Goyal, N. (2022). A novel stacking-based deterministic ensemble model for infectious disease prediction. Mathematics, 10(10), Article 1714. https://doi.org/10.3390/math10101714

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Melchane, S., Elmir, Y., & Kacimi, F. (2024). Infectious diseases prediction based on machine learning: The impact of data reduction using feature extraction techniques. Procedia Computer Science, 239, 675–683. https://doi.org/10.1016/j.procs.2024.06.223

Mirugwe, A., Ashaba, C., Namale, A., Akello, E., Bichetero, E., Kansiime, E., & Nyirenda, J. (2024). Sentiment analysis

of social media data on Ebola outbreak using deep learning classifiers. Life, 14(6), Article 708. https://doi.org/10.3390/life14060708

Moore, S., Hill, E. M., Tildesley, M. J., Dyson, L., & Keeling, M. J. (2021). Vaccination and non-pharmaceutical interventions for COVID-19: A mathematical modelling study. The Lancet Infectious Diseases, 21(6), 793–802. https://doi.org/10.1016/S1473-3099(21)00143-2

Muhammed, F. O., Suleiman, M. A., Abdullahi, S. E., & Ogar, A. O. (2025). Machine learning for epidemic outbreak prediction: Recent advances and open problems. NIPES Journal of Science and Technology Research, 7(4), 642–648. https://doi.org/10.37933/nipes/7.4.2025.SI75

Pramod, A., Abhishek, J. S., & Suganthi, K. (2023). Epidemic outbreak prediction using machine learning models. arXiv. https://doi.org/10.48550/arXiv.2310.19760

Ren, H., Ling, Y., Cao, R., Wang, Z., Li, Y., & Huang, T. (2023). Early warning of emerging infectious diseases based on multimodal data. Biosafety and Health, 5(4), 193–203. https://doi.org/10.1016/j.bsheal.2023.05.006

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Roy, P. K., & Kumar, A. (2022). Early prediction of COVID-19 using ensemble of transfer learning. Computers & Electrical Engineering, 101, Article 108018. https://doi.org/10.1016/j.compeleceng.2022.108018

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

Sebastianelli, A., Spiller, D., Carmo, R., Wheeler, J., Bonafilia, D., Gava, V., & Cox, S. (2024). A reproducible ensemble machine learning approach to forecast dengue outbreaks. Scientific Reports, 14, Article 3807. https://doi.org/10.1038/s41598-024-52796-9

Sharma, S., Gupta, Y. K., & Mishra, A. K. (2023). Analysis and prediction of COVID-19 multivariate data using deep ensemble learning methods. International Journal of Environmental Research and Public Health, 20(11), Article 5943. https://doi.org/10.3390/ijerph20115943

Shashank, P. S. T., Vaibhavi, A., Vaishnavi, J., & Jabbar, M. (2021). Regression model for prediction of epidemic outbreaks. IOP Conference Series: Materials Science and Engineering, 1042(1), Article 012016. https://doi.org/10.1088/1757-899x/1042/1/012016

Sherratt, K., Gruson, H., Grah, R., Johnson, H., Niehus, R., Prasse, B., Sandmann, F., Deuschel, J., Wolffram, D., Abbott, S., Funk, S., et al. (2023). Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. eLife, 12, Article e81916. https://doi.org/10.7554/eLife.81916

Talib, M. A., Afadar, Y., Nasir, Q., Nassif, A. B., Hijazi, H., & Hasasneh, A. (2024). A tree-based explainable AI model for early detection of COVID-19 using physiological data. BMC Medical Informatics and Decision Making, 24, Article 179. https://doi.org/10.1186/s12911-024-02576-2

Tsao, S. F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., & Butt, Z. A. (2021). What social media told us in the time of COVID-19: A scoping review. The Lancet Digital Health, 3(3), e175–e194. https://doi.org/10.1016/S2589-7500(20)30315-0

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. Artificial Intelligence Review, 53(8), 5929–5955. https://doi.org/10.1007/s10462-020-09838-1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Viana, J., van Dorp, C. H., Nunes, A., Gomes, M. C., van Boven, M., Kretzschmar, M. E., Veldhoen, M., & Rozhnova, G. (2021). Controlling the pandemic during the SARS-CoV-2 vaccination rollout. Nature Communications, 12, Article 3674. https://doi.org/10.1038/s41467-021-23938-8

Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-attention with linear complexity. arXiv. https://arxiv.org/abs/2006.04768