
Certificate Programme in AI for Public Health Surveillance

Artificial Intelligence Foundations for Public Health

Artificial Intelligence (AI) – A broad field encompassing computational techniques that enable machines to mimic human cognition. Related terms: Machine Learning, Deep Learning, Natural Language Processing. AI systems process large volumes of health data to identify patterns, predict disease spread, and support decision-making. Example: An AI model predicts influenza peaks using historical case counts and weather data. Challenges: Data quality, model transparency, and ethical considerations such as privacy and bias.

Algorithmic Bias – Systematic errors that cause an AI model to produce prejudiced outcomes against certain groups. Related terms: Fairness, Discrimination, Data Imbalance. In public health, biased algorithms may under-detect outbreaks in underserved communities. Example: A predictive model trained on urban hospital data fails to flag rural disease clusters. Challenges: Identifying hidden biases, correcting imbalanced datasets, and ensuring equitable health interventions.

Anomaly Detection – Techniques that identify data points deviating markedly from expected patterns. Related terms: Outlier Detection, Change Point Detection, Surveillance Alerts. Used in syndromic surveillance to flag unusual spikes in symptom reports. Example: A sudden rise in emergency department visits for respiratory distress triggers an alert. Challenges: Distinguishing true public health events from noise, handling seasonal variations, and setting appropriate thresholds.

Automated Text Mining – The use of algorithms to extract structured information from unstructured textual sources such as news articles, social media, or clinical notes. Related terms: Natural Language Processing, Entity Recognition, Sentiment Analysis. Enables rapid identification of emerging health threats from web streams. Example: Mining Twitter for mentions of “fever” and “cough” to map a possible COVID-19 flare-up. Challenges: Language variability, sarcasm detection, and ensuring data privacy.

Bayesian Inference – A statistical framework that updates the probability of a hypothesis as new evidence becomes available. Related terms: Prior Distribution, Posterior Probability, Markov Chain Monte Carlo. In disease modeling, Bayesian methods incorporate prior knowledge about transmission rates. Example: Updating the estimated reproduction number (R_0) as daily case counts arrive. Challenges: Choosing appropriate priors, computational intensity, and communicating uncertainty to policymakers.

Bias-Variance Trade-off – The balance between a model’s ability to capture underlying patterns (low bias) and its sensitivity to random fluctuations in training data (low variance). Related terms: Overfitting, Underfitting, Model Complexity. Selecting the right regularization technique is crucial for reliable public-health predictions. Example: A highly complex neural network fits past outbreak data perfectly but performs poorly on future seasons. Challenges: Diagnosing the source of error and adjusting model

capacity accordingly.

Case-Based Reasoning – An AI approach that solves new problems by adapting solutions from previously solved cases. Related terms: Nearest Neighbor, Instance-Based Learning, Knowledge Retrieval. Useful for rapid response when a novel disease resembles a known pathogen. Example: Comparing a new respiratory syndrome to past SARS cases to infer likely transmission routes. Challenges: Maintaining a comprehensive case repository and ensuring relevance of past cases.

Convolutional Neural Network (CNN) – A deep learning architecture specialized for processing grid-like data, such as images. Related terms: Filters, Pooling, Feature Maps. In public health, CNNs analyze radiological images to detect tuberculosis or COVID-19 pneumonia. Example: A CNN classifies chest X-rays with >90% accuracy, supporting triage in low-resource settings. Challenges: Need for large annotated datasets, interpretability of learned features, and computational resources.

Cross-Validation – A resampling method for evaluating model performance by partitioning data into training and testing subsets multiple times. Related terms: K-Fold, Hold-Out, Model Assessment. Ensures that predictive models for disease incidence are robust across different temporal or geographic splits. Example: Five-fold cross-validation of a random-forest model predicting dengue cases across provinces. Challenges: Maintaining temporal integrity (e.G., Not training on future data) and handling imbalanced outcome classes.

Data Governance – Policies, procedures, and standards that ensure data is managed responsibly throughout its lifecycle. Related terms: Data Stewardship, Privacy, Compliance. Critical for public-health surveillance systems that aggregate personal health records. Example: Implementing role-based access controls for a national disease-reporting platform. Challenges: Balancing data availability for research with legal and ethical constraints, especially under differing jurisdictional regulations.

Data Imbalance – Situations where the number of observations in one class vastly exceeds those in another, common in rare-disease surveillance. Related terms: Oversampling, Undersampling, Synthetic Minority Over-sampling Technique (SMOTE). Imbalance can cause predictive models to ignore minority events. Example: Using SMOTE to augment training data for Ebola case detection. Challenges: Avoiding over-fitting synthetic samples and preserving the true distribution of rare events.

Data Integration – The process of combining data from disparate sources into a unified view. Related terms: Data Fusion, ETL (Extract-Transform-Load), Interoperability. Enables comprehensive surveillance by linking laboratory reports, electronic health records, and mobility data. Example: Merging hospital admission logs with sewage-monitoring results to improve early detection of viral outbreaks. Challenges: Heterogeneous formats, differing terminologies, and real-time synchronization.

Decision Support System (DSS) – Computer-based tools that assist public-health officials in making informed choices by presenting analyzed data and recommendations. Related terms: Dashboard, Alert

System, Scenario Modeling. AI-enhanced DSS can simulate the impact of vaccination strategies. Example: A DSS suggests optimal allocation of limited COVID-19 vaccine doses across districts based on projected infection curves. Challenges: Ensuring usability, avoiding information overload, and maintaining up-to-date models.

Deep Learning – A subset of machine learning that employs multi-layered neural networks to automatically learn hierarchical representations from raw data. Related terms: Neural Networks, Representation Learning, Backpropagation. In public health, deep models process high-dimensional data such as genomic sequences or satellite imagery. Example: A deep recurrent network predicts malaria incidence using climate, land-use, and travel patterns. Challenges: Large computational demands, lack of transparency, and requirement for extensive labeled datasets.

Dimensionality Reduction – Techniques that compress high-dimensional data into a lower-dimensional space while preserving essential information. Related terms: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Autoencoders. Facilitates visualization and reduces overfitting in models using many health indicators. Example: Applying PCA to reduce 200 environmental variables to 10 principal components before feeding them into a predictive model. Challenges: Interpreting reduced dimensions and selecting the appropriate number of components.

Ensemble Learning – Combining predictions from multiple models to improve overall performance. Related terms: Bagging, Boosting, Stacking. Frequently used in outbreak forecasting to capture diverse patterns. Example: An ensemble of random forests, gradient-boosted trees, and support vector machines predicts weekly influenza cases with higher accuracy than any single model. Challenges: Increased computational cost, risk of correlated errors, and difficulty in interpreting the final ensemble decision.

Explainable AI (XAI) – Methods that make the inner workings of AI models understandable to human users. Related terms: Interpretability, Model Transparency, Feature Importance. In public health, XAI builds trust among clinicians and policymakers. Example: Using SHAP values to show that air-quality indices contributed most to a predicted rise in asthma exacerbations. Challenges: Balancing explanation depth with model performance and presenting insights in a non-technical manner.

Feature Engineering – The process of creating, selecting, and transforming variables to improve model performance. Related terms: Variable Construction, Encoding, Scaling. In disease surveillance, engineered features may include lagged case counts, mobility indices, or vaccination coverage rates. Example: Adding a 7-day moving average of reported symptoms as a predictor for next-day outbreak risk. Challenges: Avoiding leakage of future information, ensuring reproducibility, and handling high-frequency time-series data.

Geospatial Analysis – Techniques that incorporate geographic location into the examination of health data. Related terms: GIS (Geographic Information Systems), Spatial Autocorrelation, Heatmaps. Enables identification of disease hotspots and resource allocation. Example: Mapping dengue cases to pinpoint

neighborhoods with vector-breeding sites for targeted insecticide spraying. Challenges: Spatial data privacy, varying resolution of geographic layers, and accounting for population density differences.

Gradient Boosting Machine (GBM) – An ensemble learning method that builds models sequentially, each correcting errors of its predecessor. Related terms: XGBoost, LightGBM, AdaBoost. GBMs often achieve state-of-the-art performance in predictive public-health tasks. Example: Using XGBoost to forecast weekly COVID-19 hospital admissions based on case counts, vaccination rates, and mobility metrics. Challenges: Hyper-parameter tuning, risk of overfitting, and computational intensity on large datasets.

Health Informatics – The interdisciplinary study of the design, development, and application of information technology in health care. Related terms: Electronic Health Records (EHR), Clinical Decision Support, Interoperability. Forms the backbone for data collection in AI-driven surveillance. Example: Integrating EHR data streams with a national reporting system to enable near-real-time influenza monitoring. Challenges: Standardizing terminologies (e.G., SNOMED, ICD), ensuring data security, and addressing legacy system constraints.

Hybrid Modeling – Combining mechanistic (theory-driven) models with data-driven AI approaches to leverage strengths of both. Related terms: Compartmental Models, Machine Learning, Model Fusion. Improves forecasting when epidemiological theory alone is insufficient. Example: Embedding a SEIR (Susceptible-Exposed-Infectious-Recovered) structure within a neural network that learns residual patterns from real-world case data. Challenges: Aligning model scales, calibrating parameters jointly, and preventing contradictions between components.

Imbalanced Classification – A specific case of data imbalance where the target class distribution is skewed, common in rare-event detection such as bioterrorism threats. Related terms: Cost-Sensitive Learning, Focal Loss, ROC-AUC. Adjusting loss functions or sampling strategies can improve detection rates. Example: Applying focal loss to a CNN that identifies images of disease-specific rash, increasing sensitivity for the minority class. Challenges: Selecting appropriate evaluation metrics and avoiding inflated performance due to trivial majority-class predictions.

Incident Reporting System – Digital platforms that capture and transmit information about health events, injuries, or disease occurrences. Related terms: Surveillance System, Notification, Case Definition. AI can automate triage and prioritize alerts. Example: An AI module classifies incoming reports into “high”, “medium”, or “low” priority based on symptom severity and geographic spread. Challenges: Ensuring completeness of reports, dealing with inconsistent terminology, and preventing false alarms.

K-Nearest Neighbors (KNN) – A simple, instance-based learning algorithm that classifies a new observation based on the majority class among its k closest training points. Related terms: Distance Metric, Lazy Learning, Instance Retrieval. Useful for quick baseline models in outbreak detection. Example: Classifying a region’s risk level by comparing recent case trajectories to those of similar regions. Challenges: Sensitivity to noise, computational cost for large datasets, and choosing an appropriate k .

Knowledge Graph – A network-based representation of entities (e.G., Diseases, symptoms, locations) and their relationships. Related terms: Ontology, Triple Store, Semantic Reasoning. Enables AI to reason over complex public-health concepts. Example: A knowledge graph links COVID-19 variants to vaccine efficacy data, supporting policy decisions on booster recommendations. Challenges: Curating accurate relationships, scaling graph queries, and integrating heterogeneous data sources.

Latent Variable Model – Statistical models that infer hidden (latent) factors influencing observed data. Related terms: Factor Analysis, Hidden Markov Model, Topic Modeling. In surveillance, latent variables may capture unobserved transmission dynamics. Example: A hidden Markov model estimates the true infection prevalence from noisy test-positivity rates. Challenges: Determining the correct number of latent factors and ensuring identifiability.

Logistic Regression – A classic statistical method for modeling binary outcomes, providing interpretable coefficients reflecting odds ratios. Related terms: Generalized Linear Model, Maximum Likelihood, Regularization. Frequently used as a baseline for disease risk prediction. Example: Predicting the probability of a patient developing severe COVID-19 based on age, comorbidities, and vaccination status. Challenges: Handling non-linearity, multicollinearity, and limited predictive power when relationships are complex.

Machine Learning (ML) – Algorithms that enable computers to learn patterns from data without explicit programming. Related terms: Supervised Learning, Unsupervised Learning, Reinforcement Learning. In public health, ML powers outbreak forecasting, risk stratification, and resource optimization. Example: A random-forest model predicts weekly dengue incidence using climate, vector-control, and socioeconomic variables. Challenges: Data sparsity, model interpretability, and maintaining performance across changing epidemiological contexts.

Model Drift – The degradation of a model's predictive performance over time due to changes in underlying data distributions. Related terms: Concept Drift, Temporal Validation, Retraining. Surveillance models must be monitored continuously to detect drift. Example: A COVID-19 case-forecasting model trained on pre-vaccination data underestimates cases after mass immunization campaigns. Challenges: Establishing drift detection thresholds, scheduling timely model updates, and preserving historical comparability.

Natural Language Processing (NLP) – A suite of methods for analyzing and generating human language. Related terms: Tokenization, Named Entity Recognition, Sentiment Analysis. Enables extraction of health indicators from free-text sources such as clinical notes or social media. Example: An NLP pipeline identifies mentions of "loss of smell" in electronic health records to flag possible COVID-19 cases. Challenges: Dealing with medical jargon, multilingual corpora, and privacy-preserving text handling.

Neural Network – A computational model composed of interconnected nodes (neurons) organized in layers, capable of learning complex mappings. Related terms: Activation Function, Backpropagation, Deep Architecture. Forms the basis of most modern AI tools in public health. Example: A feed-forward network predicts hospital bed occupancy from daily admission counts and community transmission rates.

Challenges: Selecting appropriate architecture, preventing overfitting, and ensuring computational efficiency.

Outbreak Detection Algorithm – Statistical or AI-based methods designed to identify the emergence of disease clusters in near real-time. Related terms: CUSUM (Cumulative Sum), EWMA (Exponentially Weighted Moving Average), Syndromic Surveillance. Example: A CUSUM algorithm flags a significant rise in respiratory-illness reports from a city's emergency departments. Challenges: Balancing sensitivity and specificity, handling reporting delays, and integrating multiple data streams.

Precision Public Health – The application of high-resolution data and analytics to tailor interventions to specific populations or locations. Related terms: Targeted Intervention, Granular Data, Micro-Stratification. AI enables identification of hyper-local risk factors. Example: Using machine-learning models to allocate mosquito-control resources to neighborhoods with the highest predicted vector density. Challenges: Data privacy at fine geographic scales, ensuring equitable access, and avoiding stigmatization.

Probabilistic Graphical Model – A framework that represents random variables and their conditional dependencies via graphs. Related terms: Bayesian Network, Markov Random Field, Inference. Useful for modeling complex causal pathways in disease transmission. Example: A Bayesian network incorporates vaccination coverage, contact patterns, and pathogen virulence to estimate outbreak probability. Challenges: Learning accurate structure from limited data and performing efficient inference on large networks.

Public Health Surveillance System – Organized collection, analysis, and dissemination of health-related data for the purpose of disease control. Related terms: Indicator-Based Surveillance, Event-Based Surveillance, Reporting Network. AI augments these systems with automated anomaly detection and predictive analytics. Example: An AI-driven platform ingests laboratory test results, hospital admissions, and pharmacy sales to generate weekly situational reports. Challenges: Ensuring data timeliness, standardizing case definitions, and maintaining system sustainability.

Random Forest – An ensemble learning method that constructs multiple decision trees and aggregates their predictions. Related terms: Bagging, Feature Importance, Out-of-Bag Error. Frequently employed for classification of disease status and variable selection. Example: A random-forest model identifies key climatic predictors for malaria incidence across multiple regions. Challenges: Interpreting large numbers of trees, handling correlated features, and tuning the number of trees for optimal performance.

Recurrent Neural Network (RNN) – A class of neural networks designed for sequential data, where connections form directed cycles allowing information persistence. Related terms: LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), Time-Series Forecasting. RNNs excel at modeling disease trajectories over time. Example: An LSTM network predicts the next 14 days of COVID-19 case counts using past daily reports and mobility metrics. Challenges: Vanishing gradients, requirement for large sequential datasets, and high computational cost.

Reinforcement Learning (RL) – A learning paradigm where an agent interacts with an environment, receiving rewards for actions that achieve desired outcomes. Related terms: Policy, Q-Learning, Exploration-Exploitation. In public health, RL can optimize allocation of limited resources such as vaccines or testing kits. Example: An RL agent learns a vaccination schedule that minimizes projected infection peaks while respecting supply constraints. Challenges: Defining appropriate reward structures, ensuring ethical decision-making, and dealing with delayed feedback.

Risk Stratification – The process of categorizing individuals or populations based on the likelihood of experiencing adverse health events. Related terms: Scoring System, Predictive Model, Triage. AI enhances stratification by integrating diverse data sources. Example: A gradient-boosted model assigns high, medium, or low risk of severe influenza to patients based on age, comorbidities, and recent exposure history. Challenges: Avoiding discrimination, updating risk scores as epidemiology evolves, and communicating risk to end-users.

Scalable Architecture – System designs that can handle increasing data volume, velocity, and variety without performance degradation. Related terms: Distributed Computing, Cloud Infrastructure, Parallel Processing. Essential for national surveillance platforms ingesting millions of records daily. Example: Deploying a Spark-based pipeline that processes streaming hospital admission data in near real-time for outbreak alerts. Challenges: Managing cost, ensuring data security across distributed nodes, and maintaining fault tolerance.

Semantic Interoperability – The ability of different information systems to exchange data with unambiguous, shared meaning. Related terms: Ontology, Standardized Vocabulary, HL7 FHIR. Critical for integrating heterogeneous public-health datasets. Example: Mapping local disease codes to the WHO ICD-10 ontology enables cross-border outbreak comparison. Challenges: Aligning multiple terminologies, handling legacy data, and achieving consensus among stakeholders.

Sentinel Surveillance – Focused monitoring of selected health facilities or populations that act as early indicators of broader trends. Related terms: Sentinel Sites, Representative Sampling, Early Warning. AI can automate analysis of sentinel data streams. Example: An AI model evaluates weekly influenza-like illness reports from sentinel clinics to forecast national epidemic curves. Challenges: Ensuring sentinel sites remain representative, dealing with limited sample size, and integrating sentinel data with other surveillance streams.

Supervised Learning – A machine-learning paradigm where models are trained on input-output pairs to learn a mapping function. Related terms: Labels, Training Set, Regression, Classification. Most predictive public-health models, such as disease incidence forecasts, rely on supervised techniques. Example: Training a support-vector machine to classify regions as “outbreak” or “non-outbreak” based on recent case trends and environmental factors. Challenges: Obtaining high-quality labeled data, preventing overfitting, and adapting models to shifting patterns.

Time-Series Forecasting – Methods that predict future values based on previously observed temporal data.

Related terms: ARIMA, Prophet, Seasonal Decomposition. Central to anticipating disease peaks and resource needs. Example: An ARIMA model projects weekly measles cases using five years of historical surveillance data. Challenges: Capturing non-linear trends, accounting for interventions (e.G., Vaccination campaigns), and handling missing observations.

Transfer Learning – Leveraging knowledge acquired from one task or domain to improve performance on a related but distinct task. Related terms: Pre-training, Fine-tuning, Domain Adaptation. Enables rapid development of models for emerging diseases by reusing weights from existing models. Example: Fine-tuning a CNN trained on chest X-rays for pneumonia to detect COVID-19-related abnormalities with limited new images. Challenges: Risk of negative transfer when source and target domains differ substantially, and deciding which layers to freeze.

Uncertainty Quantification – Techniques that assess the confidence or variability associated with model predictions. Related terms: Confidence Intervals, Predictive Distribution, Bayesian Modeling. Provides decision-makers with risk-aware information. Example: Reporting a 95% credible interval for projected dengue cases, allowing planners to prepare for worst-case scenarios. Challenges: Computational overhead, communicating uncertainty to non-technical audiences, and integrating uncertainty into policy thresholds.

Unsupervised Learning – Algorithms that infer structure from data without explicit labels, often used for clustering or dimensionality reduction. Related terms: Clustering, Anomaly Detection, Latent Dirichlet Allocation. In surveillance, unsupervised methods can discover novel disease patterns. Example: Applying K-means clustering to symptom reports to reveal previously unrecognized syndromes. Challenges: Determining the number of clusters, interpreting results, and validating findings against clinical reality.

Validation Cohort – An independent dataset used to assess the generalizability of a predictive model after training. Related terms: External Validation, Hold-out Set, Generalization Error. Ensures that AI tools perform reliably across different populations. Example: Testing a COVID-19 severity model developed in Country A on patient data from Country B. Challenges: Accessing comparable external data, handling differences in data collection practices, and avoiding over-optimistic performance estimates.

Vector-Borne Disease Modeling – Computational approaches that simulate the transmission dynamics of diseases transmitted by arthropod vectors (e.G., Mosquitoes). Related terms: Entomological Index, Climate Suitability, Agent-Based Model. AI enhances these models by incorporating high-resolution environmental data. Example: A random-forest model predicts weekly *Aedes aegypti* abundance using temperature, precipitation, and land-cover variables, feeding into a dengue transmission simulation. Challenges: Limited vector surveillance data, complex life-cycle dynamics, and uncertainty in climate forecasts.

Web Scraping – Automated extraction of information from websites, often used to collect real-time health-related data. Related terms: Crawling, Data Harvesting, API Integration. Provides additional sources for early outbreak signals. Example: Scraping online pharmacy listings for increased sales of antipyretics as a proxy for fever prevalence. Challenges: Legal restrictions, website structure changes, and ensuring data

quality.

Zero-Shot Learning – A paradigm where a model can recognize classes it has never seen during training by leveraging auxiliary information such as semantic descriptions. Related terms: Few-Shot Learning, Attribute Embedding, Transfer Learning. In public health, useful for detecting novel pathogens with limited data. Example: A model trained on known viral image patterns predicts the presence of a newly emerging virus based on its genetic similarity to known families. Challenges: Obtaining reliable auxiliary information, ensuring robustness, and preventing misclassification of truly unknown entities.